Évaluer des distances en probabilités

Ou comment savoir *comment vite* ça se rapproche

Ferdinand Rapin

Brussels Summer School of Mathematics 2026

Contents

1. Convergences
   - Quelques jeux de probabilités
   - Ressemblances
   - Comparaisons
2. Distances entre probabilités
   - Choisir sa distance
   - Distance et vitesse
   - Encore des distances
3. Un autre aspect de la convergence
4. Conclusion

Convergences

Outline

1. Convergences
   - Quelques jeux de probabilités
   - Ressemblances
   - Comparaisons
2. Distances entre probabilités
   - Choisir sa distance
   - Distance et vitesse
   - Encore des distances
3. Un autre aspect de la convergence
4. Conclusion

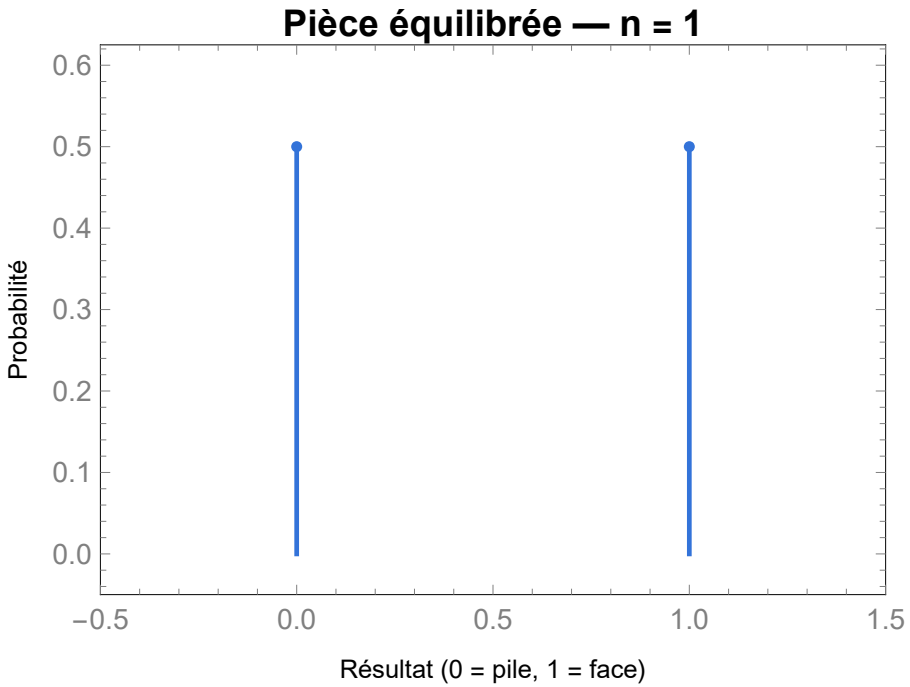
Les probabilités comme jeu

Les probabilités comme jeu

Parier sur le résultat d'un lancer de **pièce** *équilibrée*.

Les probabilités comme jeu

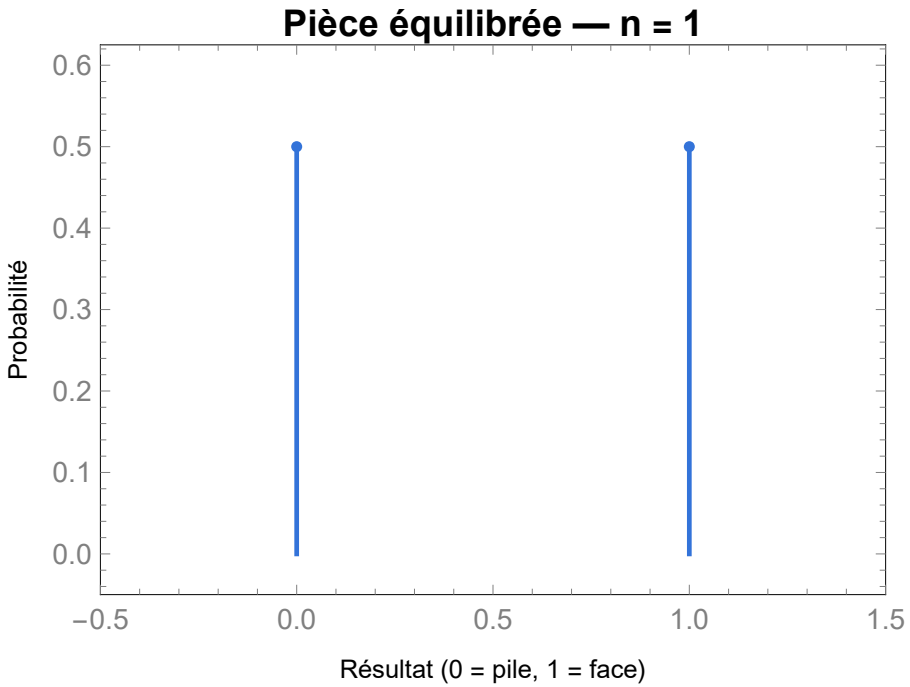
Parier sur le résultat d'un lancer de **pièce** *équilibrée*.



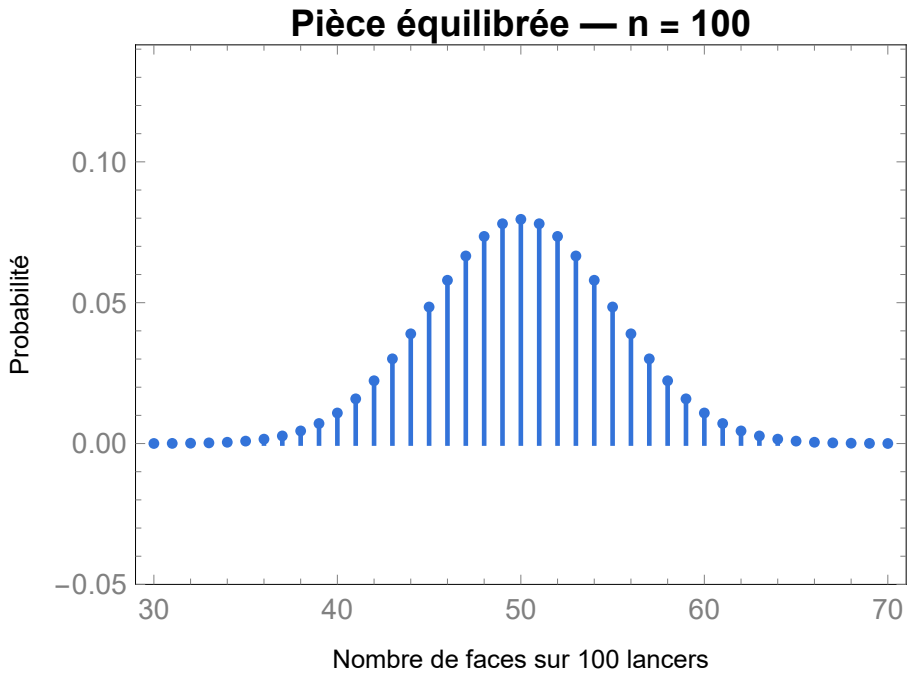
$n = 1$ partie

Les probabilités comme jeu

Parier sur le résultat d'un lancer de **pièce** *équilibrée*.



$n = 1$ partie



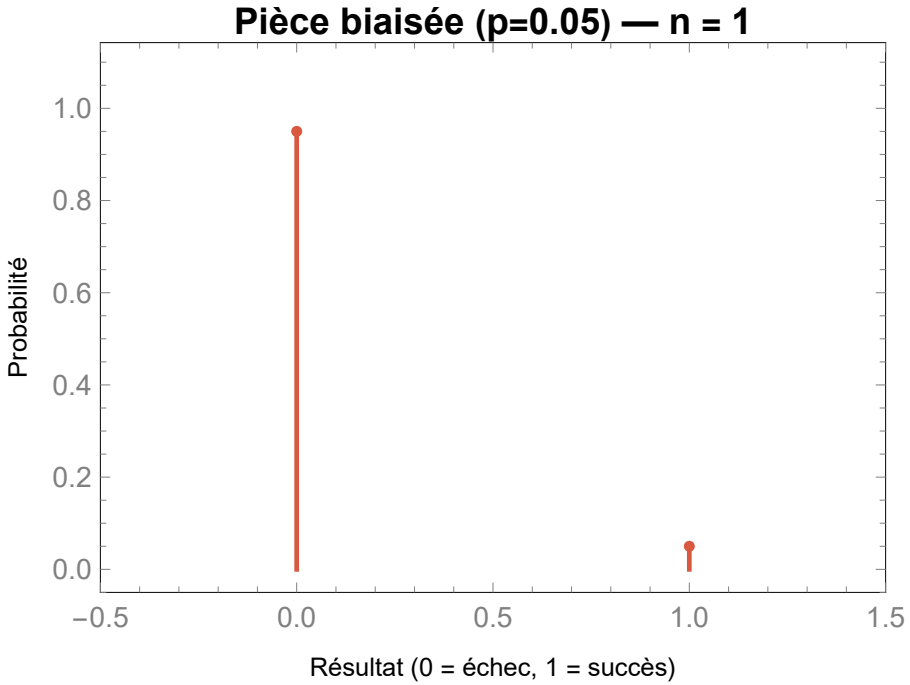
$n = 100$ parties

Les probabilités comme jeu

Parier sur le résultat d'un lancer de **pièce** *déséquilibrée* (les chances de succès sont de $p = 1/20$).

Les probabilités comme jeu

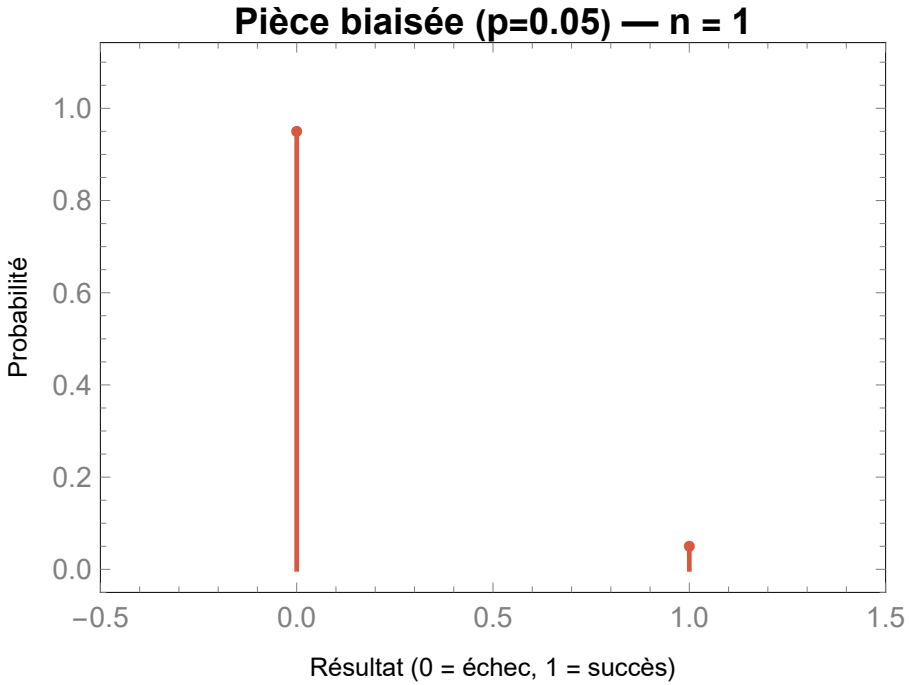
Parier sur le résultat d'un lancer de **pièce** *déséquilibrée* (les chances de succès sont de $p = 1/20$).



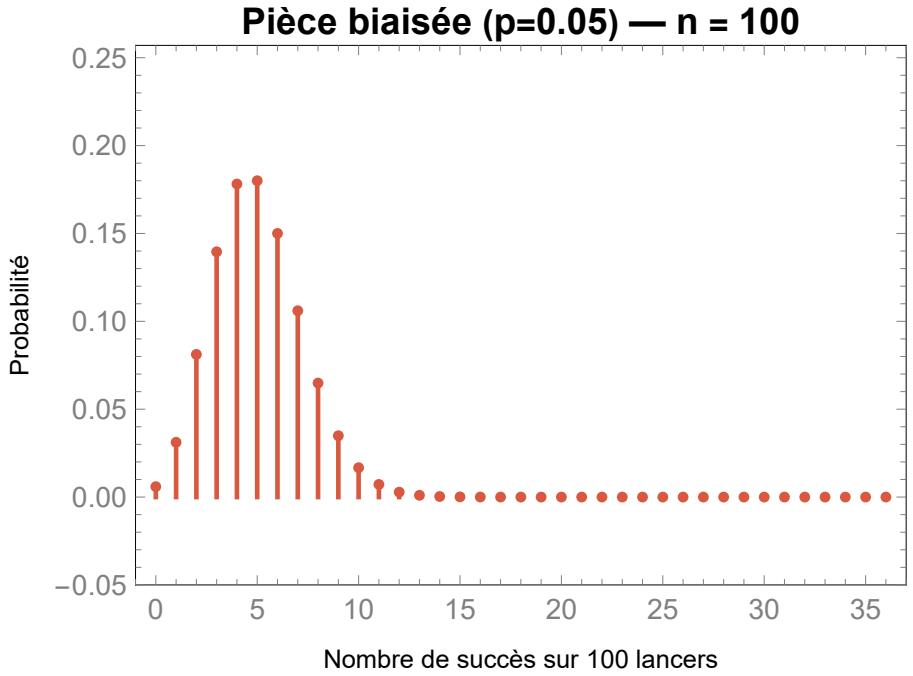
$n = 1$ partie

Les probabilités comme jeu

Parier sur le résultat d'un lancer de **pièce** *déséquilibrée* (les chances de succès sont de $p = 1/20$).



$n = 1$ partie



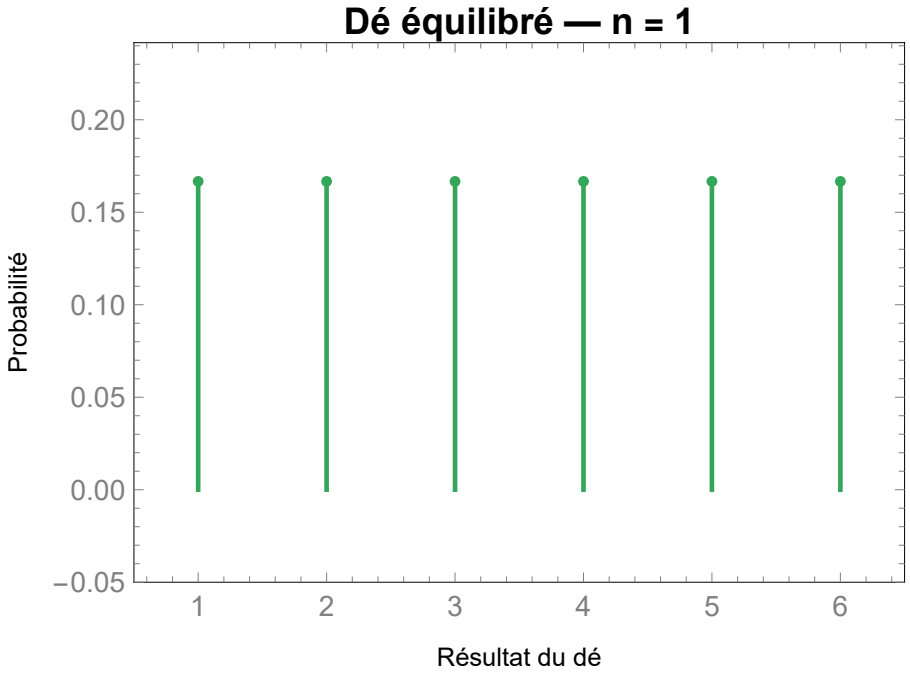
$n = 100$ parties

Les probabilités comme jeu

Parier sur le résultat d'un lancer de **dé** *équilibré*.

Les probabilités comme jeu

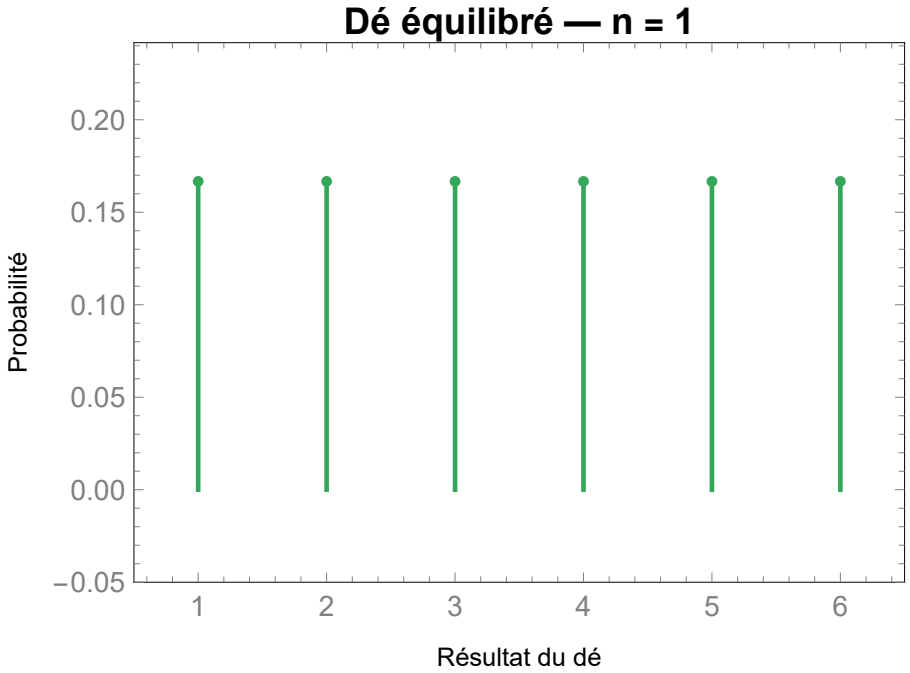
Parier sur le résultat d'un lancer de **dé** *équilibré*.



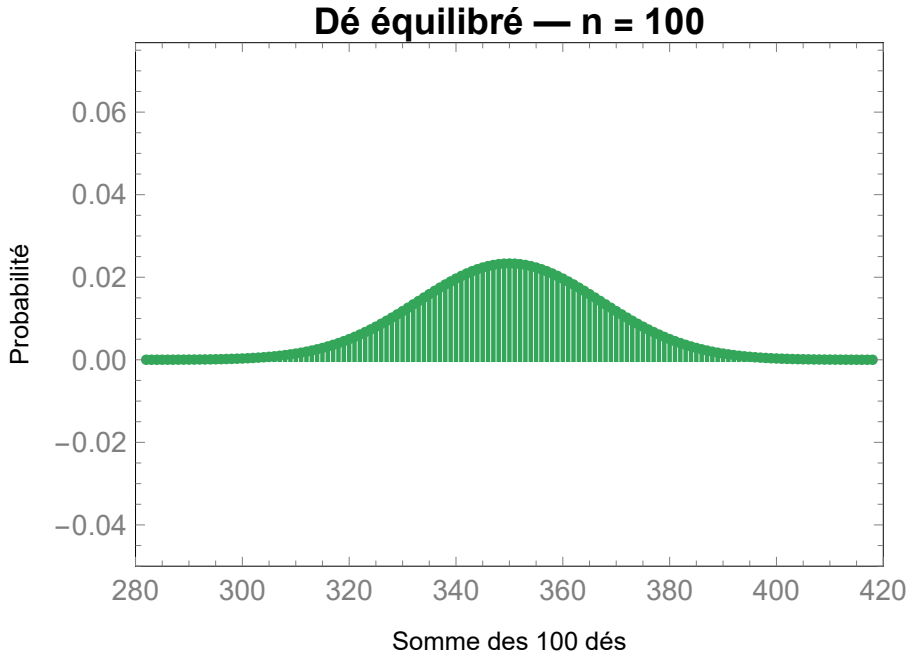
$n = 1$ partie

Les probabilités comme jeu

Parier sur le résultat d'un lancer de **dé** *équilibré*.



$n = 1$ partie



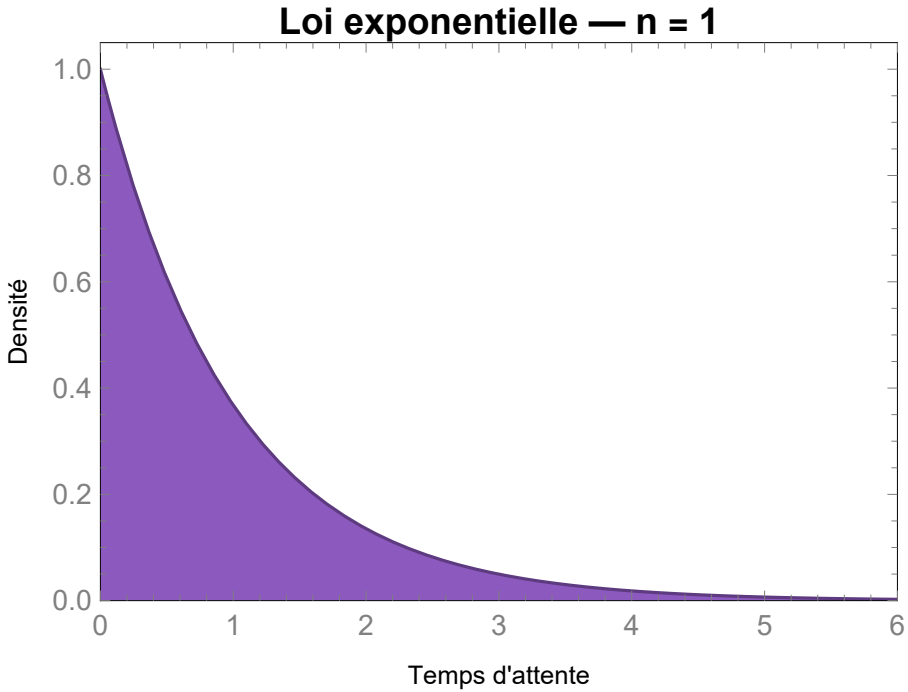
$n = 100$ parties

Les probabilités comme jeu

Regarder le temps d'attente à un arrêt de bus, entre l'entrée de deux clients, entre deux coups de téléphone, ...

Les probabilités comme jeu

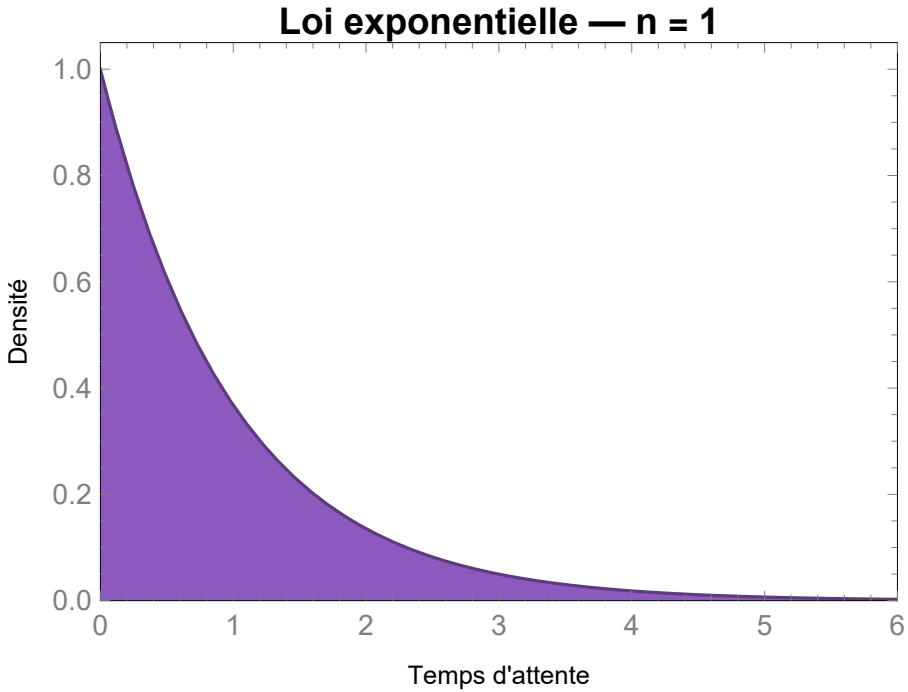
Regarder le temps d'attente à un arrêt de bus, entre l'entrée de deux clients, entre deux coups de téléphone, ...



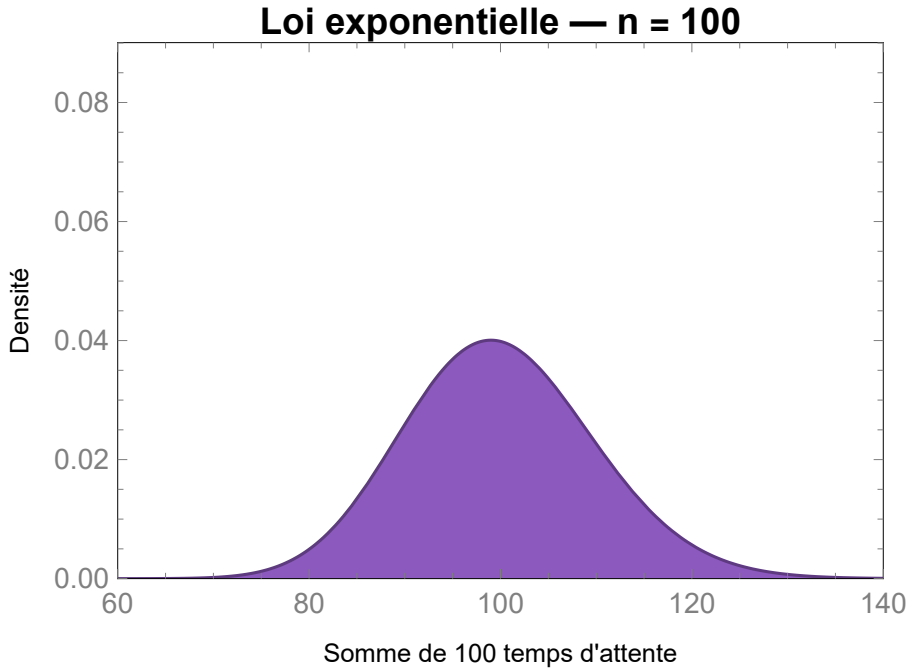
$n = 1$ partie

Les probabilités comme jeu

Regarder le temps d'attente à un arrêt de bus, entre l'entrée de deux clients, entre deux coups de téléphone, ...



$n = 1$ partie



$n = 100$ parties

Outline

1. Convergences
   - Quelques jeux de probabilités
   - Ressemblances
   - Comparaisons
2. Distances entre probabilités
   - Choisir sa distance
   - Distance et vitesse
   - Encore des distances
3. Un autre aspect de la convergence
4. Conclusion

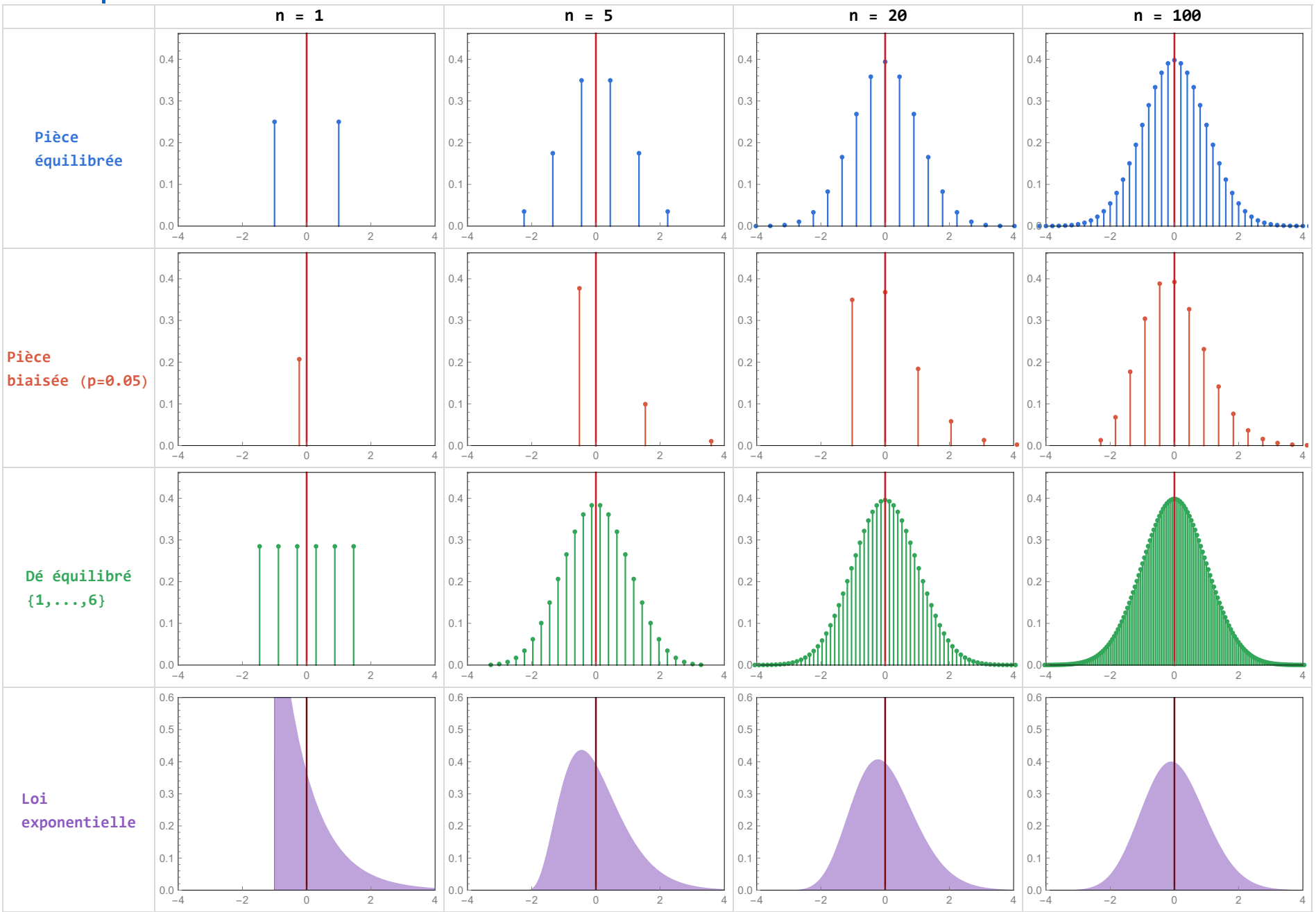
Comparons

Expérience	μ	σ	μ_{100}	σ_{100}
Pièce équilibrée $(p = 0.5)$	0.5	0.5	50	5
Pièce biaisée $(p = 0.05)$	0.05	≈ 0.218	5	≈ 2.179
Dé équilibré $\{1, \dots, 6\}$	3.5	1.708	350	≈ 17.078
Loi exponentielle(1)	1	1	100	10

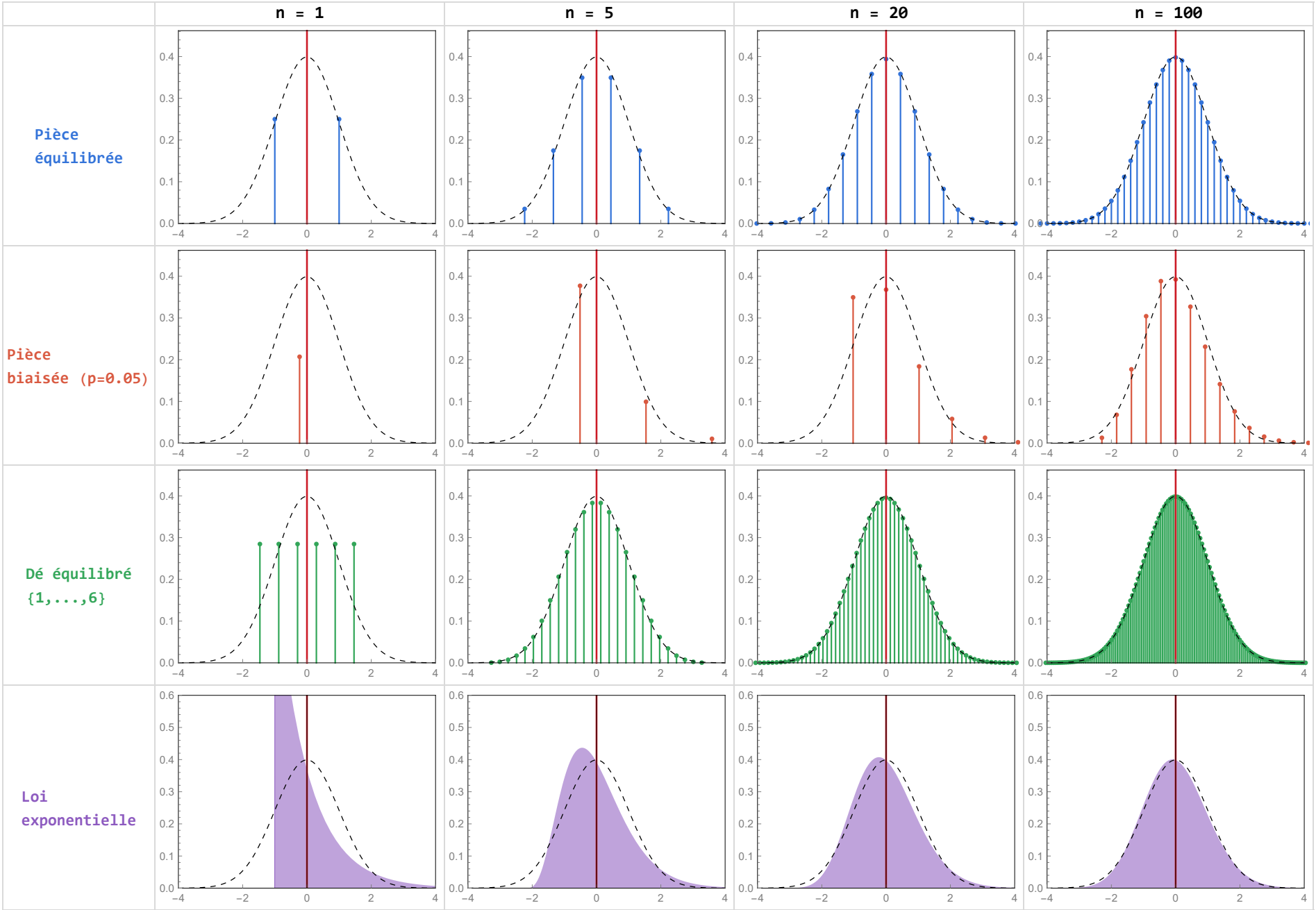
Table: Moyenne et écart-type de chaque expérience de base, et de la somme S_{100} sur $n = 100$ répétitions.

Pour pouvoir comparer, nous allons centrer *(imposer $\mu = 0$)* et réduire *(imposer $\sigma = 1$)* nos résultats. On appelle ça *standardiser*.

Comparons


n = 1
n = 5
n = 20
n = 100
Pièce équilibrée
Pièce biaisée (p=0.05)
Dé équilibré {1,...,6}
Loi exponentielle

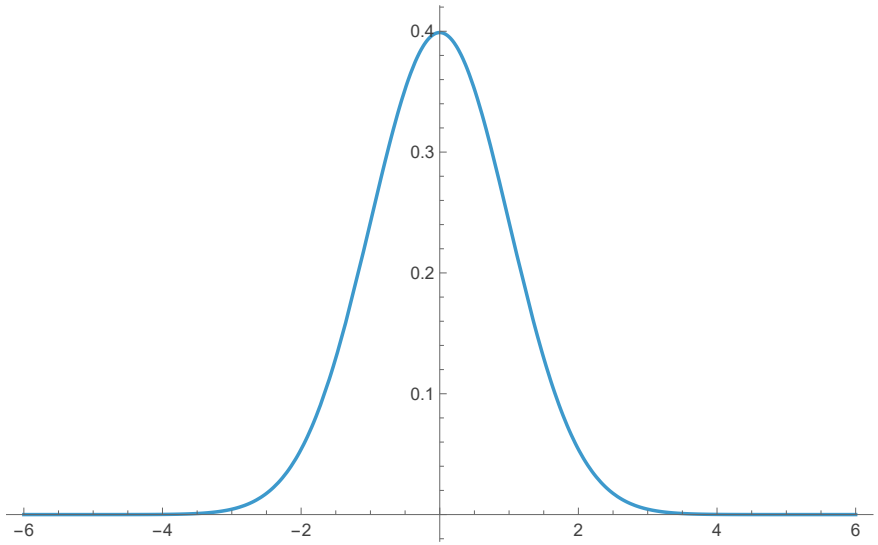
Similarités ?



La loi Gaussienne

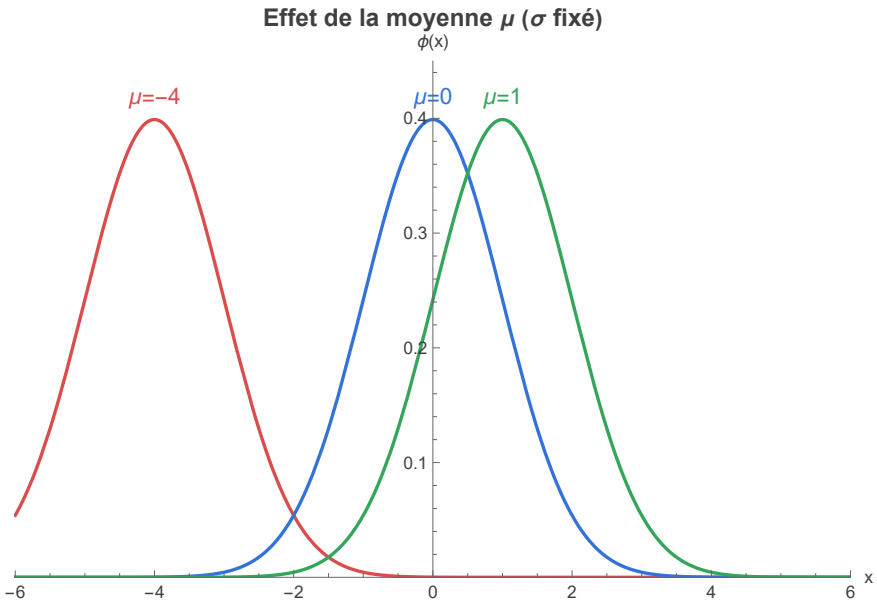
La fonction qui représente la distribution est

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

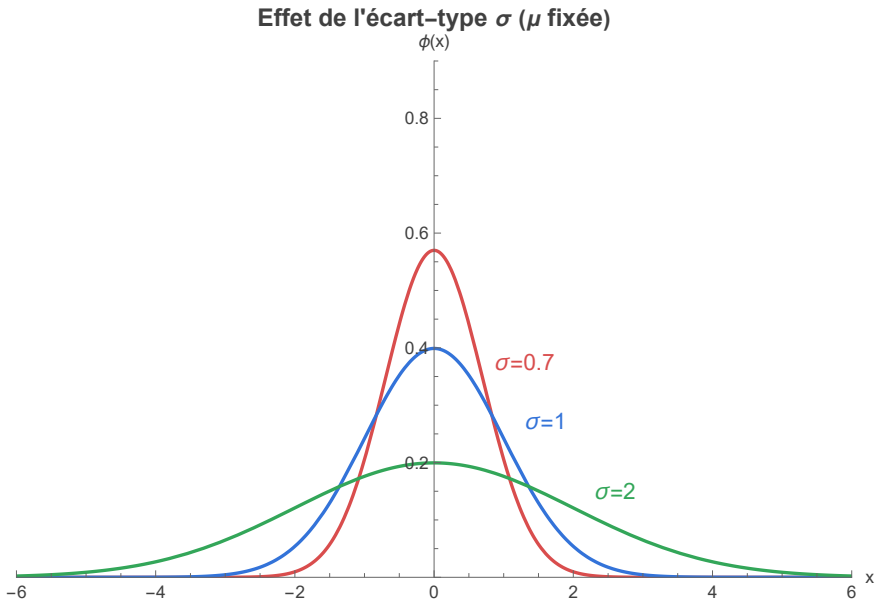


(a) Densité de la loi normale, $\mu = 0$, $\sigma = 1$

Propriétés



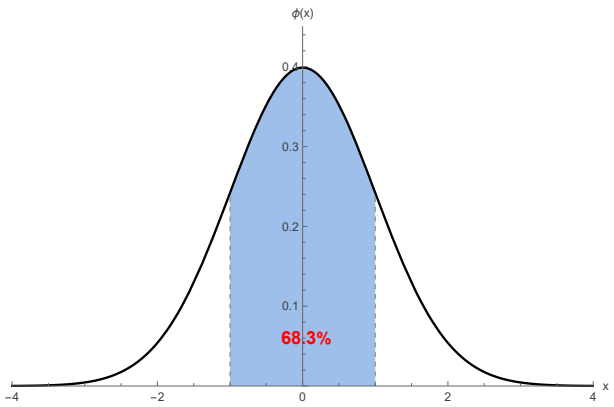
(a) Effet de la moyenne



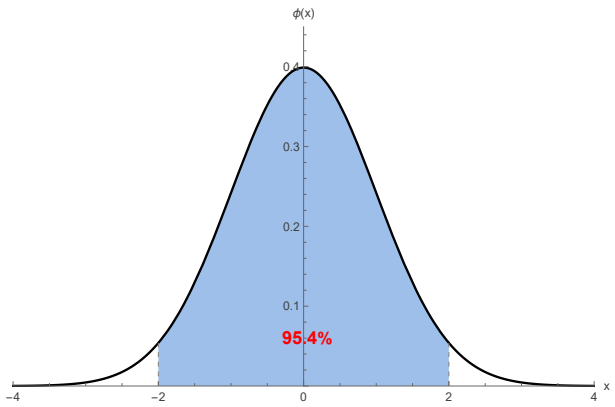
(b) Effet de l'écart-type

Figure: Densités gaussiennes pour différentes valeurs des paramètres

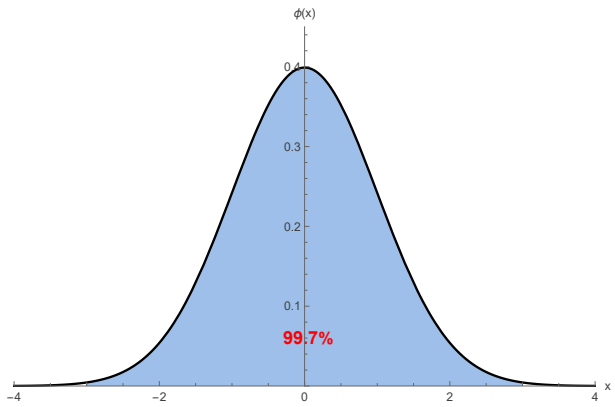
Propriétés



(a) $k = 1$



(b) $k = 2$



(c) $k = 3$

Figure: Densité gaussienne standard avec la masse comprise entre $-k\sigma$ et $k\sigma$ mise en évidence.

Les figures ci-dessus montrent que la loi gaussienne est adaptée pour notre intuition. Si une observation est "loin" de la moyenne, elle est peu probable.

Convergence !

Théorème (TCL)

Soit $X_1, X_2, \ldots$ des répétitions indépendantes d'une même expérience aléatoire, de moyenne μ et d'écart-type σ (fini). Notons

$$S_n = X_1 + X_2 + \cdots + X_n, \qquad Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Alors, lorsque $n \to \infty$,

$$Z_n \xrightarrow{\text{loi}} \mathcal{N}(0,1).$$

Convergence !

Théorème (TCL)

Soit $X_1, X_2, \ldots$ des répétitions indépendantes d'une même expérience aléatoire, de moyenne μ et d'écart-type σ (fini). Notons

$$S_n = X_1 + X_2 + \cdots + X_n, \qquad Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Alors, lorsque $n \to \infty$,

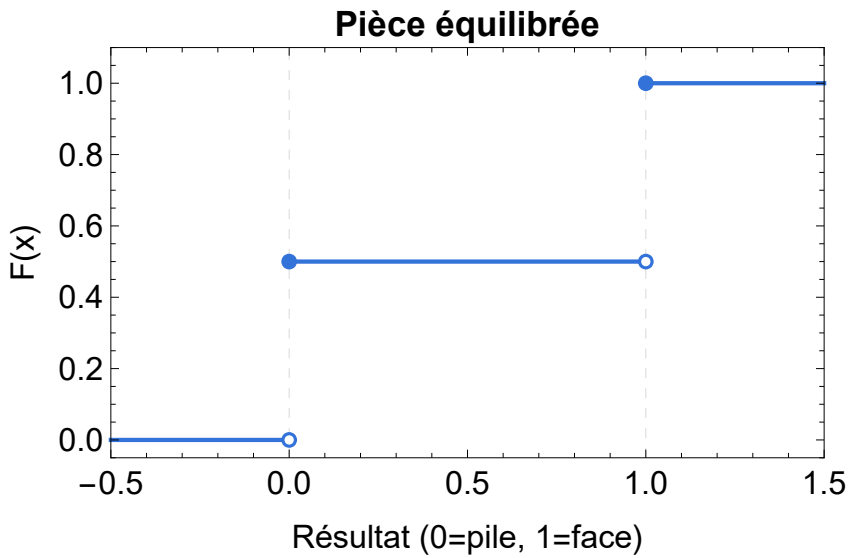
$$Z_n \xrightarrow{\text{loi}} \mathcal{N}(0,1).$$

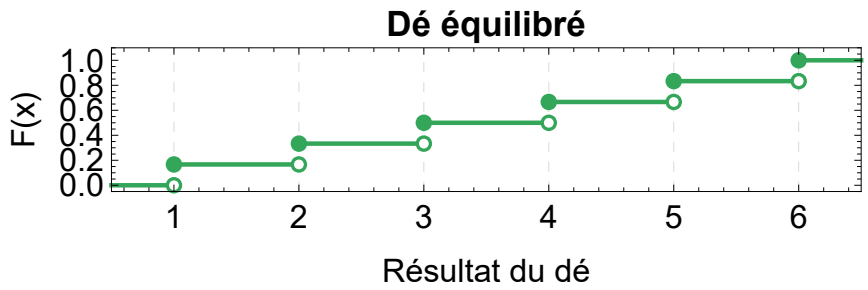
La convergence en loi est liée à la fonction de répartition, $F_X(x) = \Pr(X \leq x)$.

Fonction de répartition

Dans le cas discret,

$$F_X(x) = \Pr(X \leq x) = \sum_{y=-\infty}^{x} \Pr(X = y) = \sum_{y=-\infty}^{x} p_X(y).$$

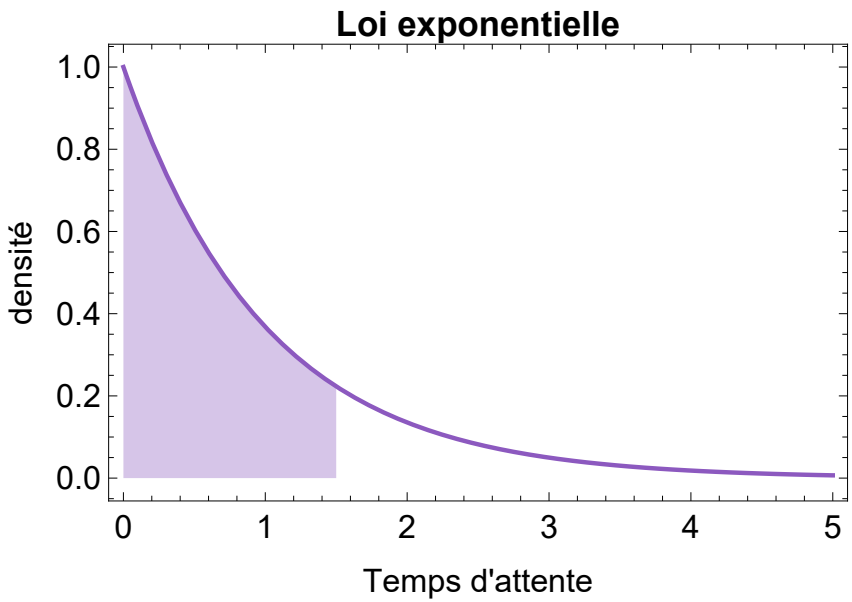


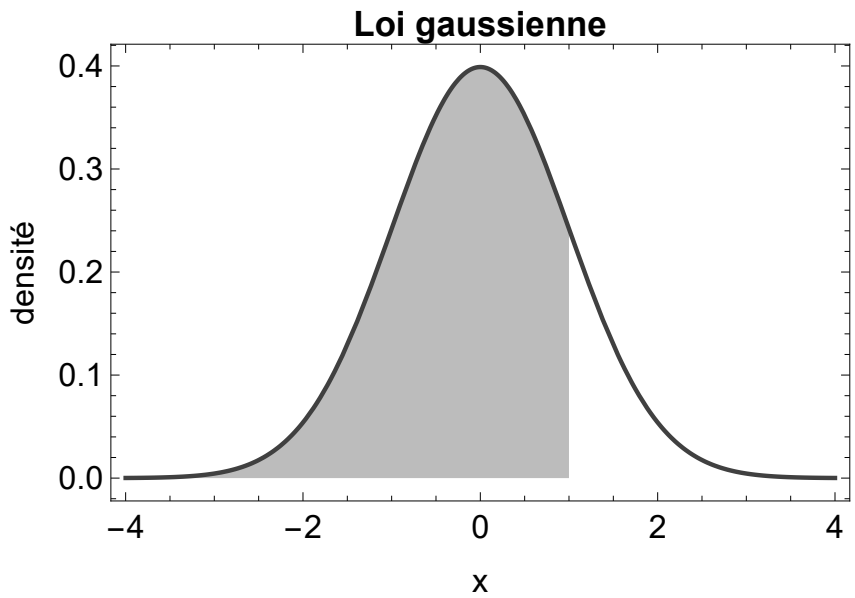


Fonction de répartition

Dans le cas continu,

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} p_X(t)dt.$$

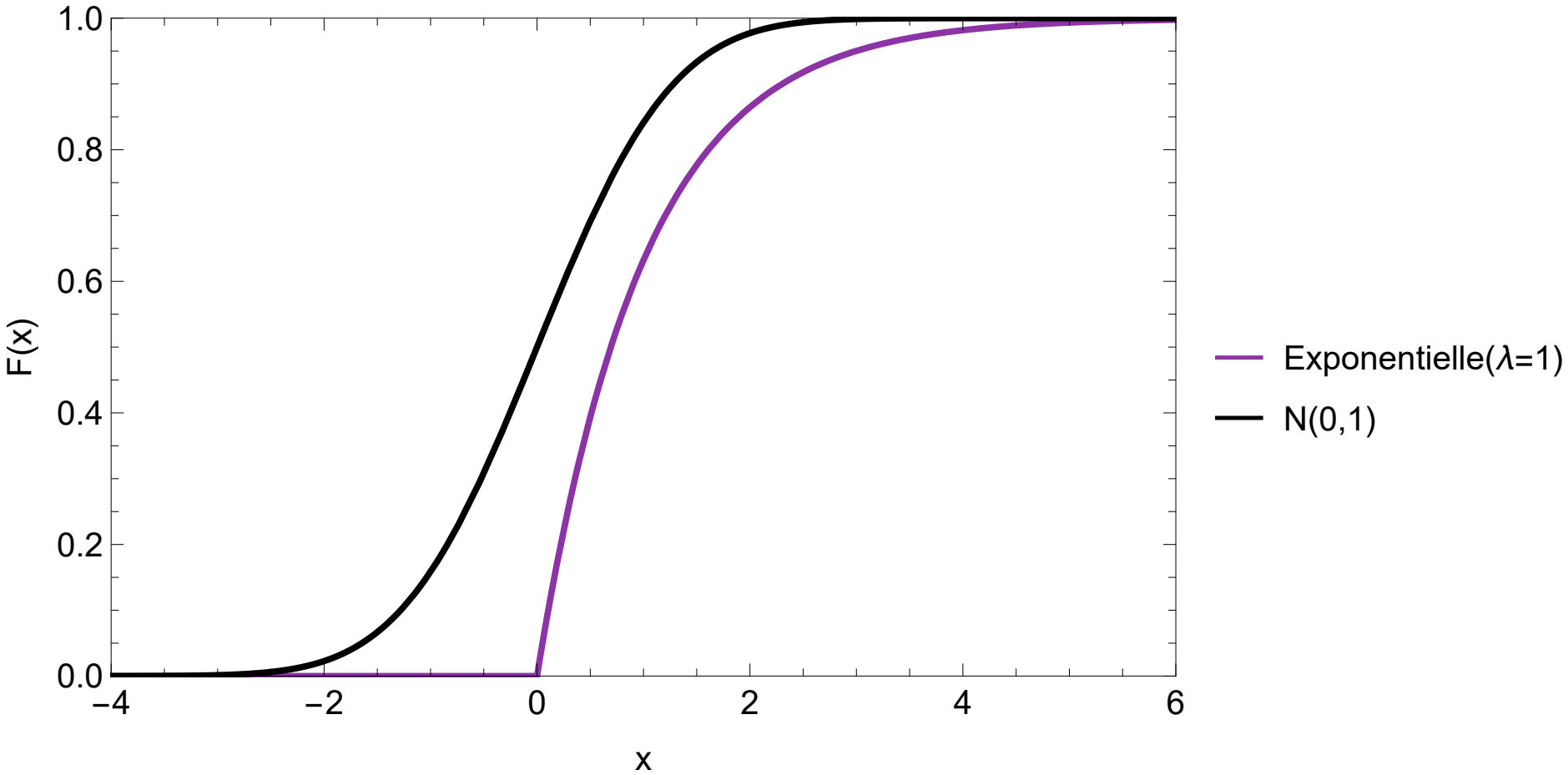




Fonction de répartition

Dans le cas continu,

$$F_X(x) = \Pr(X \leq x) = \int_{-\infty}^{x} p_X(t)dt.$$



Convergence en loi

$$X_n \xrightarrow{\text{loi}} Y \Leftrightarrow \lim_{n\to\infty} F_{X_n}(x) = F_Y(x) \quad \text{(aux point de continuité de } F_Y\text{)}.$$

Convergence en loi

$$X_n \xrightarrow{\text{loi}} Y \Leftrightarrow \lim_{n\to\infty} F_{X_n}(x) = F_Y(x) \quad \text{(aux point de continuité de } F_Y\text{)}.$$

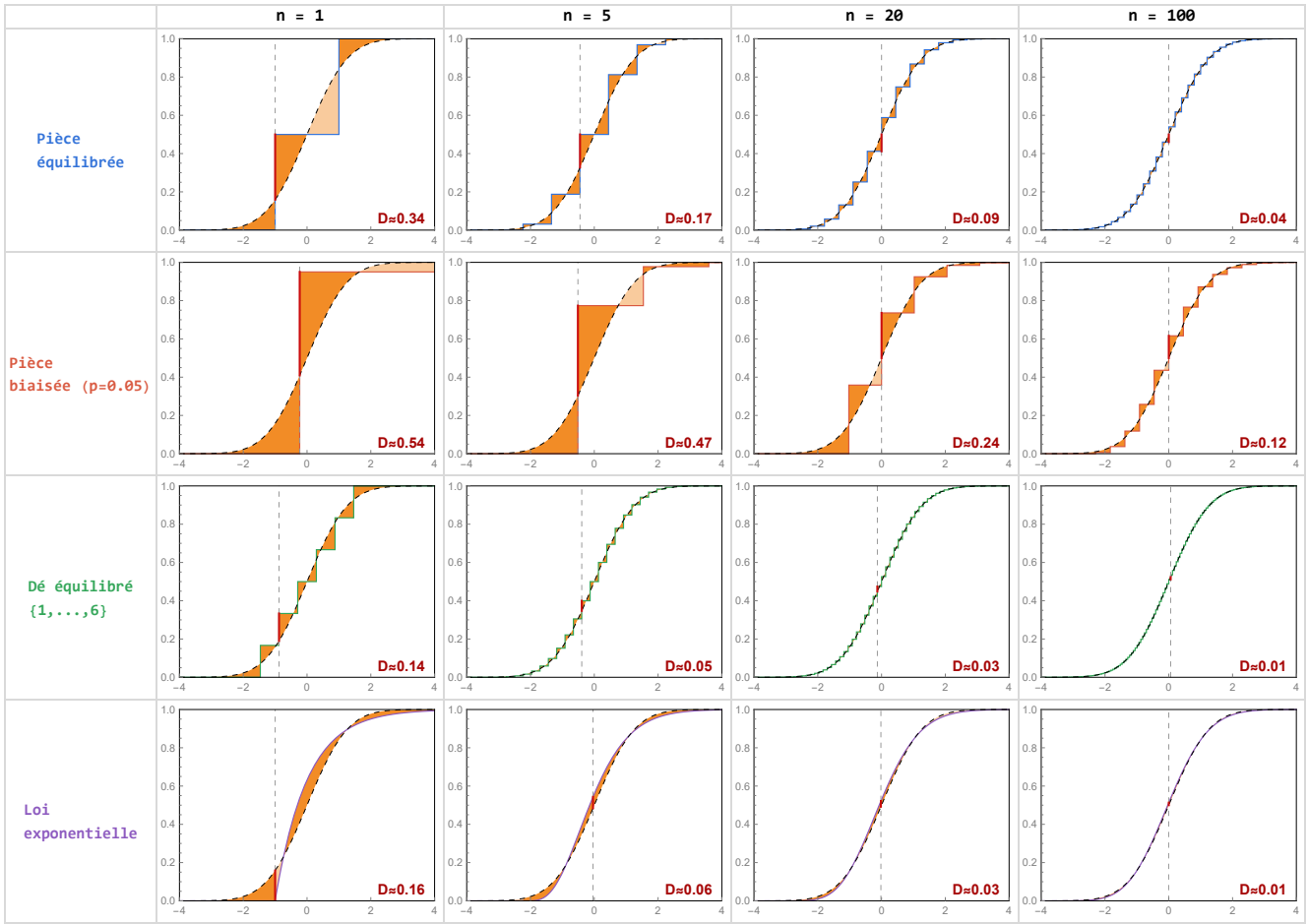


Figure: Fonctions de répartition des différents jeux

Outline

1. Convergences
   - Quelques jeux de probabilités
   - Ressemblances
   - Comparaisons
2. Distances entre probabilités
   - Choisir sa distance
   - Distance et vitesse
   - Encore des distances
3. Un autre aspect de la convergence
4. Conclusion

Vitesse de convergence : Berry–Esseen

Le TCL affirme que Z_n se rapproche de $\mathcal{N}(0,1)$, mais ne dit rien sur *la vitesse de ce rapprochement* – ni même comment mesurer un tel écart entre deux lois.

Vitesse de convergence : Berry–Esseen

Le TCL affirme que Z_n se rapproche de $\mathcal{N}(0,1)$, mais ne dit rien sur *la vitesse de ce rapprochement* – ni même comment mesurer un tel écart entre deux lois.

Au vu de la définition de la convergence en loi, il est naturel de regarder la *distance de Kolmogorov* :

$$d_{\mathrm{K}}(X,Y) = \sup_x |F_X(x) - F_Y(x)|.$$

Vitesse de convergence : Berry–Esseen

Le TCL affirme que Z_n se rapproche de $\mathcal{N}(0,1)$, mais ne dit rien sur *la vitesse de ce rapprochement* – ni même comment mesurer un tel écart entre deux lois.

Au vu de la définition de la convergence en loi, il est naturel de regarder la *distance de Kolmogorov* :

$$d_{\mathrm{K}}(X,Y) = \sup_x |F_X(x) - F_Y(x)|.$$

Théorème (Berry–Esseen)

Soit Z_n, la standardisation de la somme des X_i, et N une gaussienne. Supposons que ρ (la moyenne de $|X_1 - \mu|^3$) soit finie. Il existe alors une constante universelle C – qui ne dépend ni de n, ni de la loi de X_1 – telle que

$$d_{\mathrm{K}}(Z_n, N) \ \leq \ \frac{C\,\rho}{\sigma^3\sqrt{n}}.$$

La distance de Kolmogorov

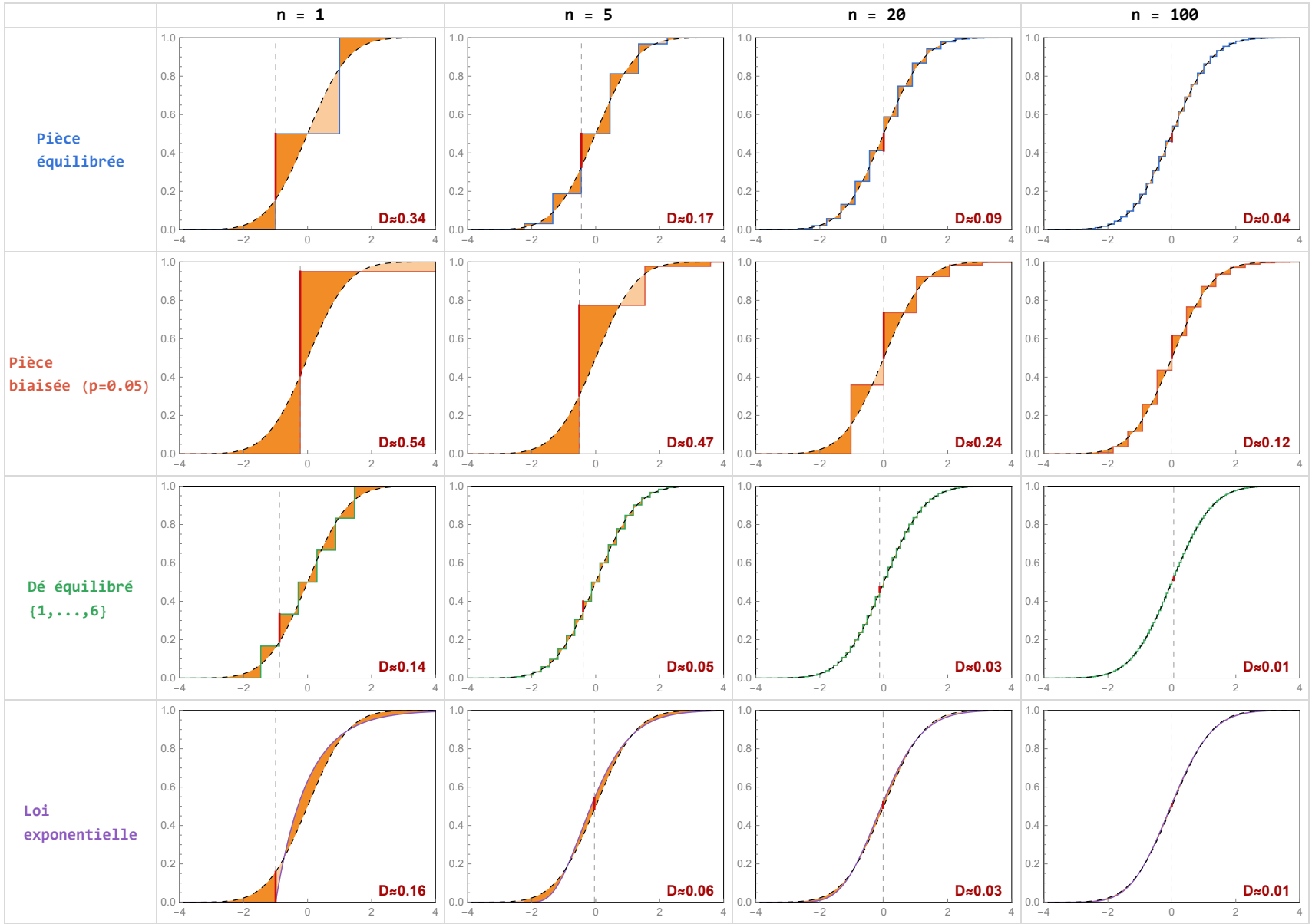


Figure: Fonctions de répartition des différents jeux et distance de Kolmogorov (rouge)

LA distance ?

La distance de Kolmogorov représente le pire désaccord en un point, naturel ?

LA distance ?

La distance de Kolmogorov représente le pire désaccord en un point, naturel ?

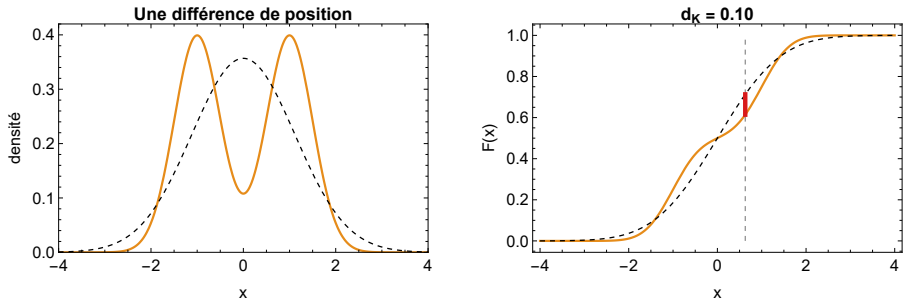


Figure: La courbe orange est un mélange de deux lois normales à deux moyenne différente (-1 et 1), d'écart-type 0.5. À gauche sa fonction de densité, à droite sa fonction de répartition

LA distance ?

La distance de Kolmogorov représente le pire désaccord en un point, naturel ?

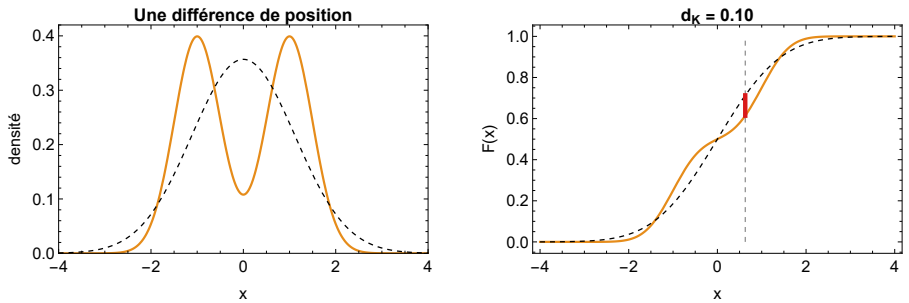


Figure: La courbe orange est un mélange de deux lois normales à deux moyenne différente (-1 et 1), d'écart-type 0.5. À gauche sa fonction de densité, à droite sa fonction de répartition

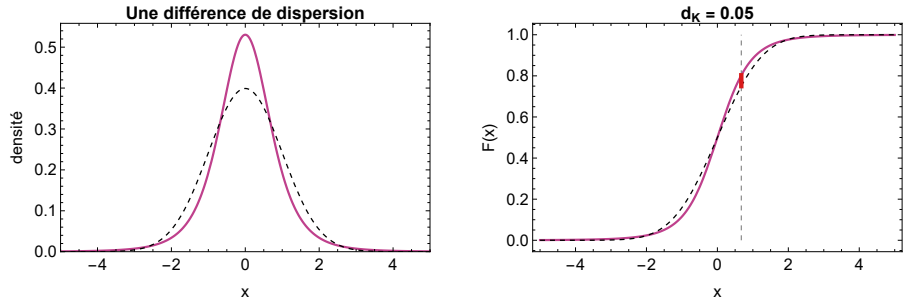


Figure: La courbe rose est une distribution de Student de paramètre 4, qui a la propriété d'avoir plus de poids aux extrêmes que la loi normale. À gauche sa fonction de densité, à droite sa fonction de répartition

Distances entre probabilités

D'autres distances - 1

La particularité de la distance de Kolmogorov : regarder en un seul point x,

$$d_{\mathrm{K}}(X, Y) = \sup_x |\Pr(X \leq x) - \Pr(Y \leq x)|.$$

D'autres distances - 1

La particularité de la distance de Kolmogorov : regarder en un seul point x,

$$d_{\mathrm{K}}(X, Y) = \sup_x |\Pr(X \leq x) - \Pr(Y \leq x)|.$$

Une alternative naturelle : regarder à plusieurs endroits à la fois, compter le pire écart sur une "zone" (*un ensemble mesurable)*.

D'autres distances - 1

La particularité de la distance de Kolmogorov : regarder en un seul point x,

$$d_{\mathrm{K}}(X, Y) = \sup_x |\Pr(X \leq x) - \Pr(Y \leq x)|.$$

Une alternative naturelle : regarder à plusieurs endroits à la fois, compter le pire écart sur une "zone" (*un ensemble mesurable)*.

On obtient alors la *distance en variation totale*

$$d_{\mathrm{TV}}(X, Y) = \sup_B |\Pr(X \in B) - \Pr(Y \in B)|.$$

Variation totale - Exemple

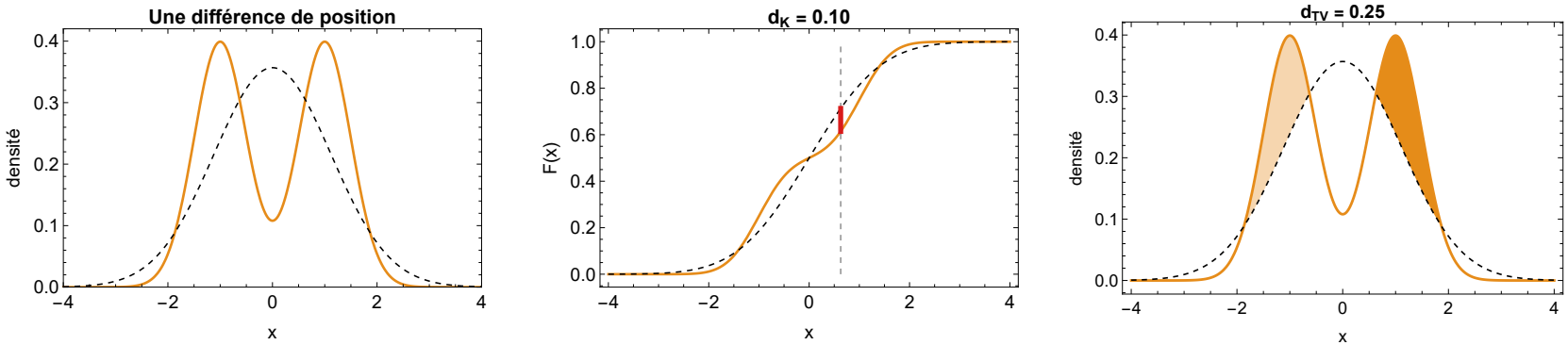


Figure: La courbe orange est la somme de deux gaussiennes. À gauche sa fonction de densité, au milieu, le calcule de la distance de Kolmogorov, à droite le calcule de la distance en variation totale.

Variation totale - Exemple

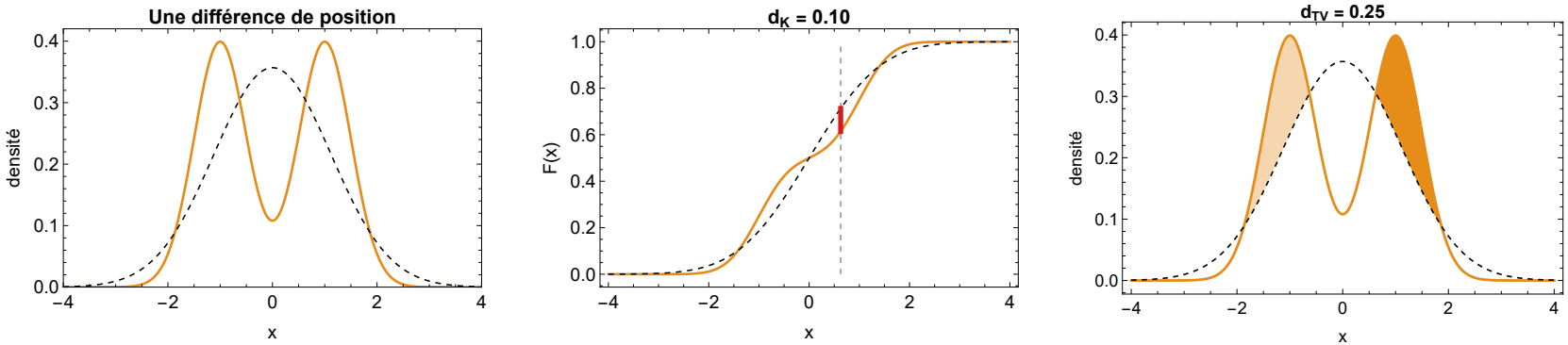


Figure: La courbe orange est la somme de deux gaussiennes. À gauche sa fonction de densité, au milieu, le calcule de la distance de Kolmogorov, à droite le calcule de la distance en variation totale.

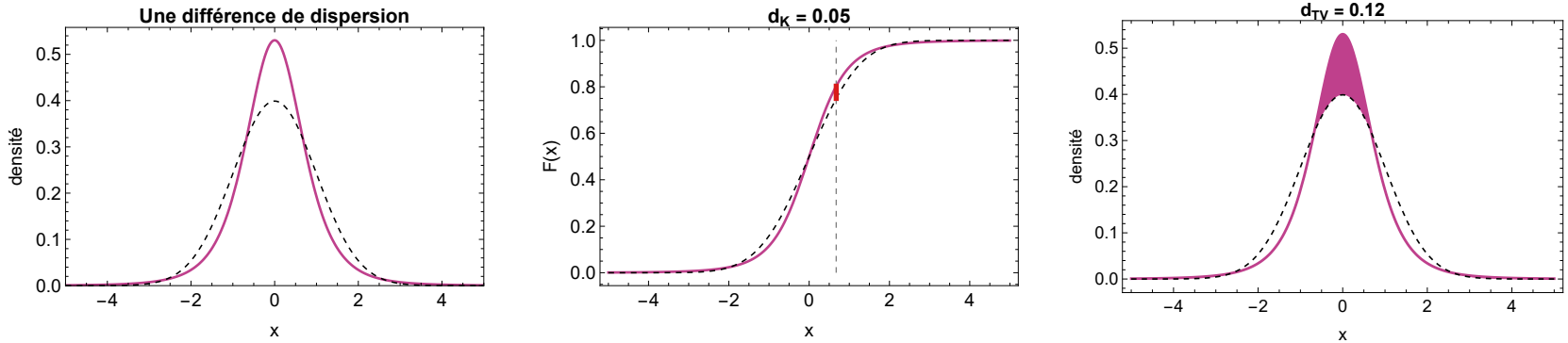


Figure: La courbe rose est une distribution de Student de paramètre 4. À gauche sa fonction de densité, au milieu, le calcule de la distance de Kolmogorov, à droite le calcule de la distance en variation totale.

D'autres distances - 2

La particularité de la distance de Kolmogorov : regarder en un seul point x :

$$d_{\mathrm{K}}(X, Y) = \sup_x |F_X(x) - F_Y(x)|.$$

D'autres distances - 2

La particularité de la distance de Kolmogorov : regarder en un seul point x :

$$d_{\mathrm{K}}(X, Y) = \sup_x |F_X(x) - F_Y(x)|.$$

Une alternative naturelle (encore ?) : compter toutes les différences ponctuelles.

D'autres distances - 2

La particularité de la distance de Kolmogorov : regarder en un seul point x :

$$d_{\mathrm{K}}(X, Y) = \sup_x |F_X(x) - F_Y(x)|.$$

Une alternative naturelle (encore ?) : compter toutes les différences ponctuelles.

On obtient alors la *distance de Wasserstein (d'ordre 1)*

$$d_{\mathrm{W}_1}(X, Y) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)| \mathrm{d}t.$$

Wasserstein - Exemple

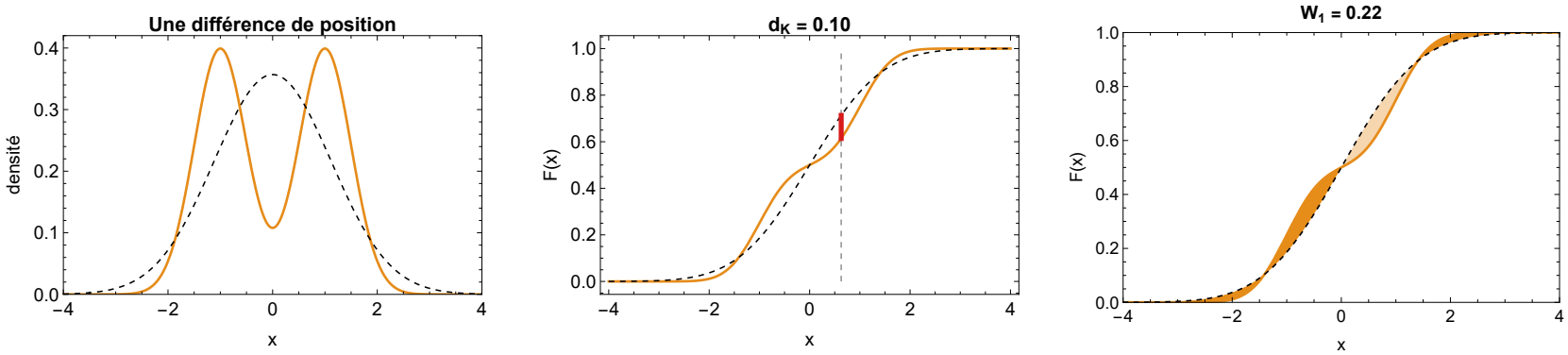


Figure: La courbe orange est la somme de deux gaussiennes. À droite sa fonction de densité, au milieu, le calcul de la distance de Kolmogorov, à droite le calcule de la distance de Wasserstein-1.

Wasserstein - Exemple

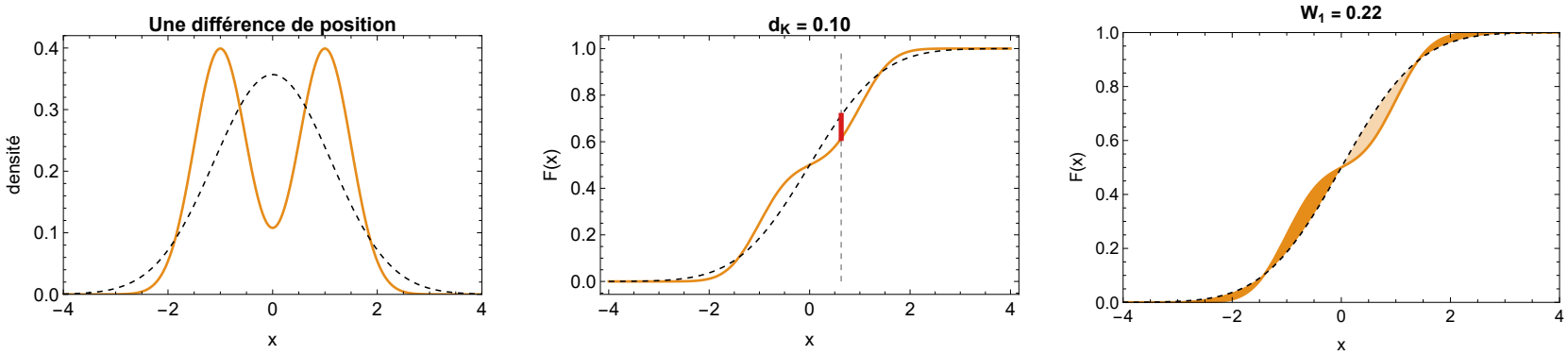


Figure: La courbe orange est la somme de deux gaussiennes. À droite sa fonction de densité, au milieu, le calcul de la distance de Kolmogorov, à droite le calcule de la distance de Wasserstein-1.

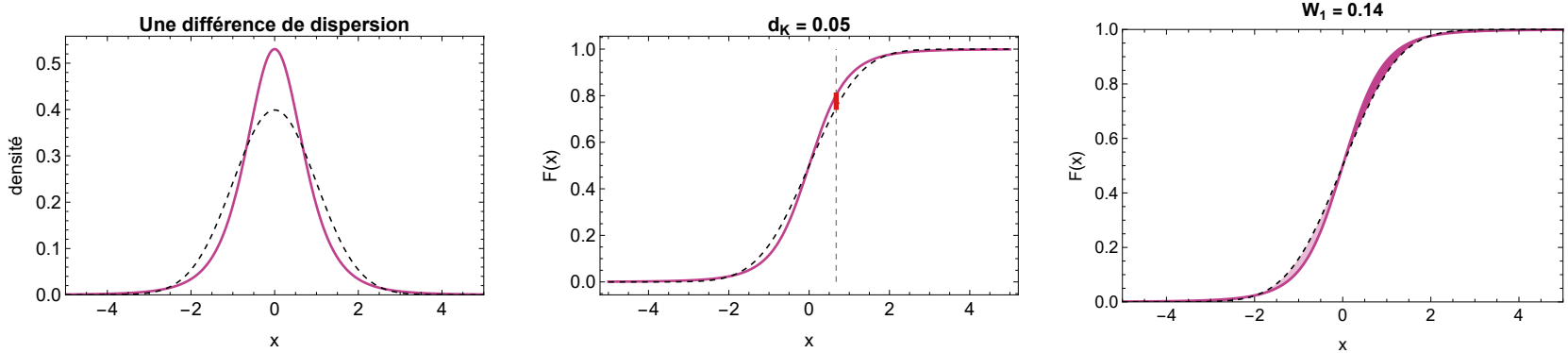


Figure: La courbe rose est une distribution de Student de paramètre 4. À droite sa fonction de densité, au milieu, le calcul de la distance de Kolmogorov, à droite le calcule de la distance de Wasserstein-1.

Outline

1. Convergences
   - Quelques jeux de probabilités
   - Ressemblances
   - Comparaisons
2. Distances entre probabilités
   - Choisir sa distance
   - Distance et vitesse
   - Encore des distances
3. Un autre aspect de la convergence
4. Conclusion

Une distance = une question

Chacune de ces distances *teste* un type de question sur les deux lois.

Une distance = une question

Chacune de ces distances *teste* un type de question sur les deux lois.

- Kolmogorov et Variation totale : "où est positionnée la masse" (parmi les intervalles $(-\infty, x]$ ou parmi les ensembles mesurables)

Une distance = une question

Chacune de ces distances *teste* un type de question sur les deux lois.

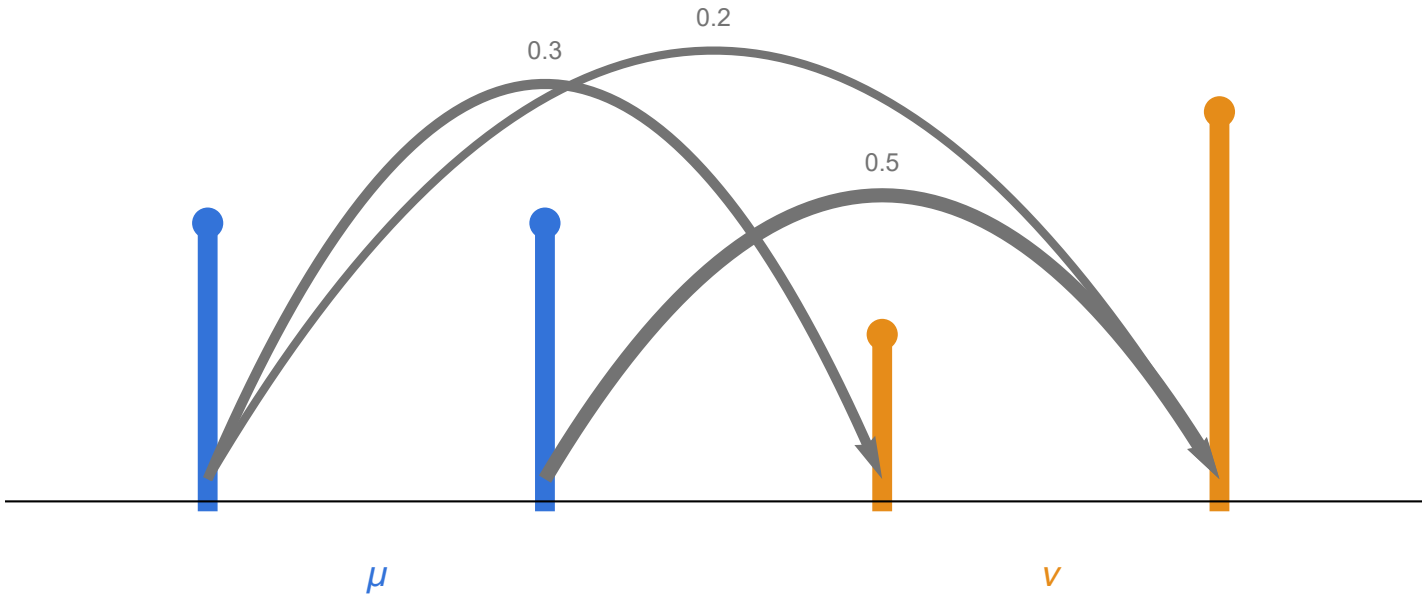
- Kolmogorov et Variation totale : "où est positionnée la masse" (parmi les intervalles $(-\infty, x]$ ou parmi les ensembles mesurables)
- Kolmogorov et Wasserstein : "quelle est l'ampleur des variations ponctuelles ?" (l'unique pire, ou bien l'ampleur totale)

Une distance = une question

Chacune de ces distances *teste* un type de question sur les deux lois.

- Kolmogorov et Variation totale : "où est positionnée la masse" (parmi les intervalles $(-\infty, x]$ ou parmi les ensembles mesurables)
- Kolmogorov et Wasserstein : "quelle est l'ampleur des variations ponctuelles ?" (l'unique pire, ou bien l'ampleur totale)
- Variation totale et Wasserstein :

Le problème du transport optimal



$$0{,}3 \cdot 2 + 0{,}2 \cdot 3 + 0{,}5 \cdot 2 = 2{,}2$$

Une distance = une question

Chacune de ces distances *teste* un type de question sur les deux lois.

1. Kolmogorov et Variation totale : "où est positionnée la masse" (parmi les intervalles $(-\infty, x]$ ou parmi les ensembles mesurables)
2. Kolmogorov et Wasserstein : "quelle est l'ampleur des variations ponctuelles ?" (l'unique pire, ou bien l'ampleur totale)
3. Variation totale et Wasserstein : "quel est le coût du transport optimal entre les deux densités ?"

Une distance = une question

Chacune de ces distances *teste* un type de question sur les deux lois.

1. Kolmogorov et Variation totale : "où est positionnée la masse" (parmi les intervalles $(-\infty, x]$ ou parmi les ensembles mesurables)
2. Kolmogorov et Wasserstein : "quelle est l'ampleur des variations ponctuelles ?" (l'unique pire, ou bien l'ampleur totale)
3. Variation totale et Wasserstein : "quel est le coût du transport optimal entre les deux densités ?" (avec une fonction de coût $c(x, y) = \mathbf{1}_{x \neq y}$ ou $c(x, y) = |x - y|$).

Si toutes les distances s'interprètent comme un même mécanisme (poser une question, regarder l'écart de réponse), on doit pouvoir décrire ce mécanisme mathématiquement !

Comment poser une question à la loi

Une loi de probabilité ne s'observe jamais directement, c'est une règle abstraite.

Comment poser une question à la loi

Une loi de probabilité ne s'observe jamais directement, c'est une règle abstraite. Tout ce qu'on peut en apprendre passe par une seule opération : lui **poser une question** h (une fonction), et lire sa **réponse** $\mathbb{E}[h(X)]$.

Le symbole $\mathbb{E}$ est celui de *l'opérateur* « espérance » et s'interprète comme la "réponse attendue" dans le cadre de la loi, à une certaine fonction-question.

Comment poser une question à la loi

Une loi de probabilité ne s'observe jamais directement, c'est une règle abstraite. Tout ce qu'on peut en apprendre passe par une seule opération : lui **poser une question** h (une fonction), et lire sa **réponse** $\mathbb{E}[h(X)]$.

Le symbole $\mathbb{E}$ est celui de *l'opérateur* « espérance » et s'interprète comme la "réponse attendue" dans le cadre de la loi, à une certaine fonction-question.

Comment ça fonctionne ?

Comment poser une question à la loi

Une loi de probabilité ne s'observe jamais directement, c'est une règle abstraite. Tout ce qu'on peut en apprendre passe par une seule opération : lui **poser une question** h (une fonction), et lire sa **réponse** $\mathbb{E}[h(X)]$.

Le symbole $\mathbb{E}$ est celui de *l'opérateur* « espérance » et s'interprète comme la "réponse attendue" dans le cadre de la loi, à une certaine fonction-question.

Comment ça fonctionne ?

1. Par sa densité, la loi "associe" un poids $p(x)$ à "chaque élément" x,

Comment poser une question à la loi

Une loi de probabilité ne s'observe jamais directement, c'est une règle abstraite. Tout ce qu'on peut en apprendre passe par une seule opération : lui **poser une question** h (une fonction), et lire sa **réponse** $\mathbb{E}[h(X)]$.

Le symbole $\mathbb{E}$ est celui de *l'opérateur* « espérance » et s'interprète comme la "réponse attendue" dans le cadre de la loi, à une certaine fonction-question.

Comment ça fonctionne ?

1. Par sa densité, la loi "associe" un poids $p(x)$ à "chaque élément" x,
2. On associe ce poids à chaque réponse $h(x)$,

Comment poser une question à la loi

Une loi de probabilité ne s'observe jamais directement, c'est une règle abstraite. Tout ce qu'on peut en apprendre passe par une seule opération : lui **poser une question** h (une fonction), et lire sa **réponse** $\mathbb{E}[h(X)]$.

Le symbole $\mathbb{E}$ est celui de *l'opérateur* « espérance » et s'interprète comme la "réponse attendue" dans le cadre de la loi, à une certaine fonction-question.

Comment ça fonctionne ?

1. Par sa densité, la loi "associe" un poids $p(x)$ à "chaque élément" x,
2. On associe ce poids à chaque réponse $h(x)$,
3. On compte toute les contributions (on fait la "moyenne") $\rightarrow$ l'espérance.

Comment poser une question à la loi

Une loi de probabilité ne s'observe jamais directement, c'est une règle abstraite. Tout ce qu'on peut en apprendre passe par une seule opération : lui **poser une question** h (une fonction), et lire sa **réponse** $\mathbb{E}[h(X)]$.

Le symbole $\mathbb{E}$ est celui de *l'opérateur* « espérance » et s'interprète comme la "réponse attendue" dans le cadre de la loi, à une certaine fonction-question.

Comment ça fonctionne ?

1. Par sa densité, la loi "associe" un poids $p(x)$ à "chaque élément" x,
2. On associe ce poids à chaque réponse $h(x)$,
3. On compte toute les contributions (on fait la "moyenne") $\rightarrow$ l'espérance.

Formellement,

$$\mathbb{E}[h(X)] = \begin{cases} \sum_i h(x_i)p(x_i) \\ \int_{\mathbb{R}} h(x)p(x)dx \end{cases}$$

Quelles questions ?

Example (Exemple de questions)

« Quelle proportion de toi est sous t ? »	$h(x) = \mathbf{1}_{x \leq t}$
« Quelle proportion de toi est dans l'ensemble B ? »	$h(x) = \mathbf{1}_{x \in B}$
« Où es-tu centrée ? »	$h(x) = x$
« À quel point es-tu dispersée ? »	$h(x) = (x - \mathbb{E}[X])^2$
« Es-tu symétrique ? »	$h(x) = ((x - \mathbb{E}[X])/\sigma_X)^3$
« As-tu beaucoup de poids aux extrêmes »	$h(x) = ((x - \mathbb{E}[X])/\sigma_X)^4$

Et plein d'autres ...

Remarque

Si deux lois répondent exactement pareil à *toutes* les questions possibles, alors ce sont la même loi.

Des questions aux distances

La distance de Kolmogorov se réécrit

$$d_{\mathrm{K}}(X,Y) = \sup_x |F_X(x) - F_Y(x)| = \sup_x |\mathbb{E}[\mathbf{1}_{X\leq x}] - \mathbb{E}[\mathbf{1}_{Y\leq x}]|$$
$$= \sup_{h:h(\cdot)=\mathbf{1}_{\cdot\leq x}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|.$$

Des questions aux distances

La distance de Kolmogorov se réécrit

$$d_{\mathrm{K}}(X,Y) = \sup_x |F_X(x) - F_Y(x)| = \sup_x |\mathbb{E}[\mathbf{1}_{X\leq x}] - \mathbb{E}[\mathbf{1}_{Y\leq x}]|$$
$$= \sup_{h:h(\cdot)=\mathbf{1}_{\cdot\leq x}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|.$$

La distance en variation totale

$$d_{\mathrm{TV}}(X,Y) = \sup_B |\Pr(X\in B) - \Pr(Y\in B)| = \sup_{h:h(\cdot)=\mathbf{1}_{\cdot\in B}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|.$$

Des questions aux distances

La distance de Kolmogorov se réécrit

$$d_{\mathrm{K}}(X,Y) = \sup_x |F_X(x) - F_Y(x)| = \sup_x |\mathbb{E}[\mathbf{1}_{X\leq x}] - \mathbb{E}[\mathbf{1}_{Y\leq x}]|$$
$$= \sup_{h:h(\cdot)=\mathbf{1}_{\cdot\leq x}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|.$$

La distance en variation totale

$$d_{\mathrm{TV}}(X,Y) = \sup_B |\Pr(X\in B) - \Pr(Y\in B)| = \sup_{h:h(\cdot)=\mathbf{1}_{\cdot\in B}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|.$$

De manière beaucoup moins évidente

$$d_{\mathrm{W}_1}(X,Y) = \int_{\mathbb{R}} |F_X(t) - F_Y(t)|\mathrm{d}t = \sup_{h:\|h'\|\leq 1} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|$$

par dualité de Kantorovich–Rubinstein

Autre distance, autre convergence

Definition (Integral Probability metric)

Une integral Probability metric (IPM) est une "distance" entre deux lois de probabilités définie à partir d'une classe de fonctions tests $\mathcal{H}$ **séparante** :

$$d_{\mathcal{H}}(X, Y) = \sup_{h \in \mathcal{H}} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|.$$

Outline

1. Convergences
   - Quelques jeux de probabilités
   - Ressemblances
   - Comparaisons
2. Distances entre probabilités
   - Choisir sa distance
   - Distance et vitesse
   - Encore des distances
3. Un autre aspect de la convergence
4. Conclusion

Distances différentes, vitesses différentes - 1

Reprenons la pièce équilibrée du début, avec n lancers. Soit $S_n \sim \mathrm{Bin}(n, p)$ et

$$Z_n = \frac{S_n - np}{\sqrt{np(1-p)}}, \qquad N \sim \mathcal{N}(0, 1).$$

Z_n est la standardisation de S_n. Le théorème central limite dit $Z_n \xrightarrow{\text{loi}} N$.

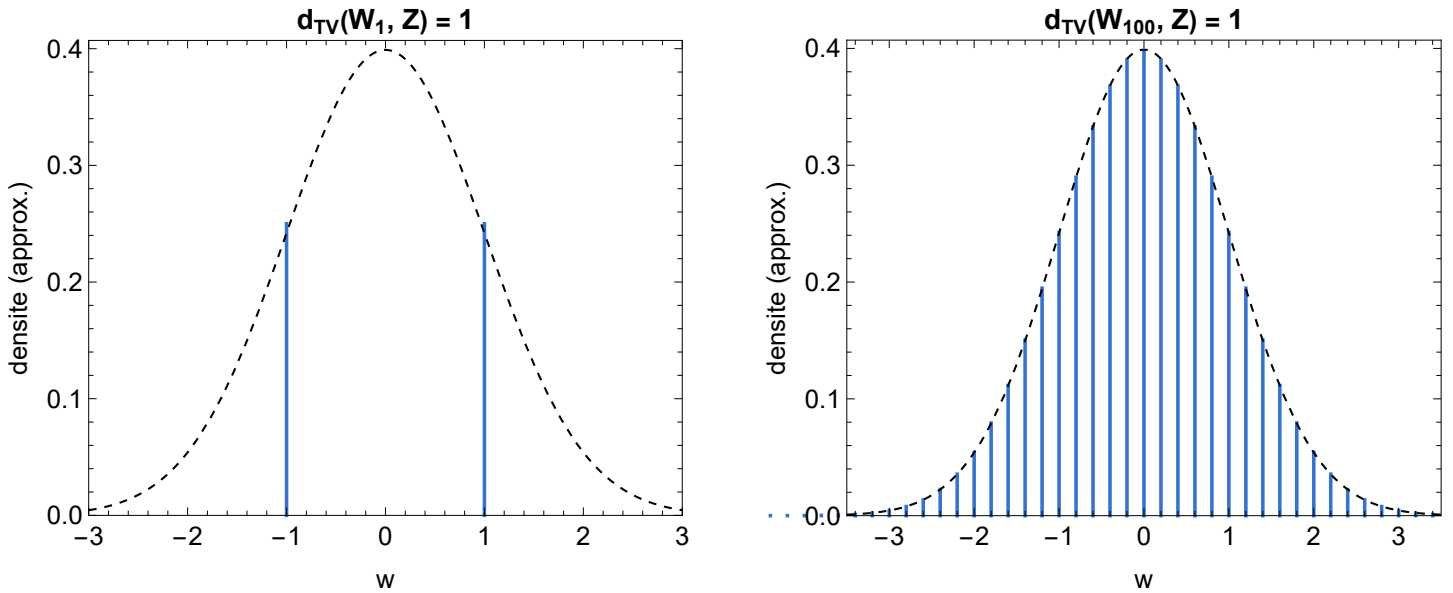


Figure: Fonction de masse du score standardisé Z_n de l'expérience "lancer n pièce", pour $n = 1$ (gauche) et $n = 100$ (droite). La courbe pointillée est la densité gaussienne.

Distances différentes, vitesses différentes - 1bis

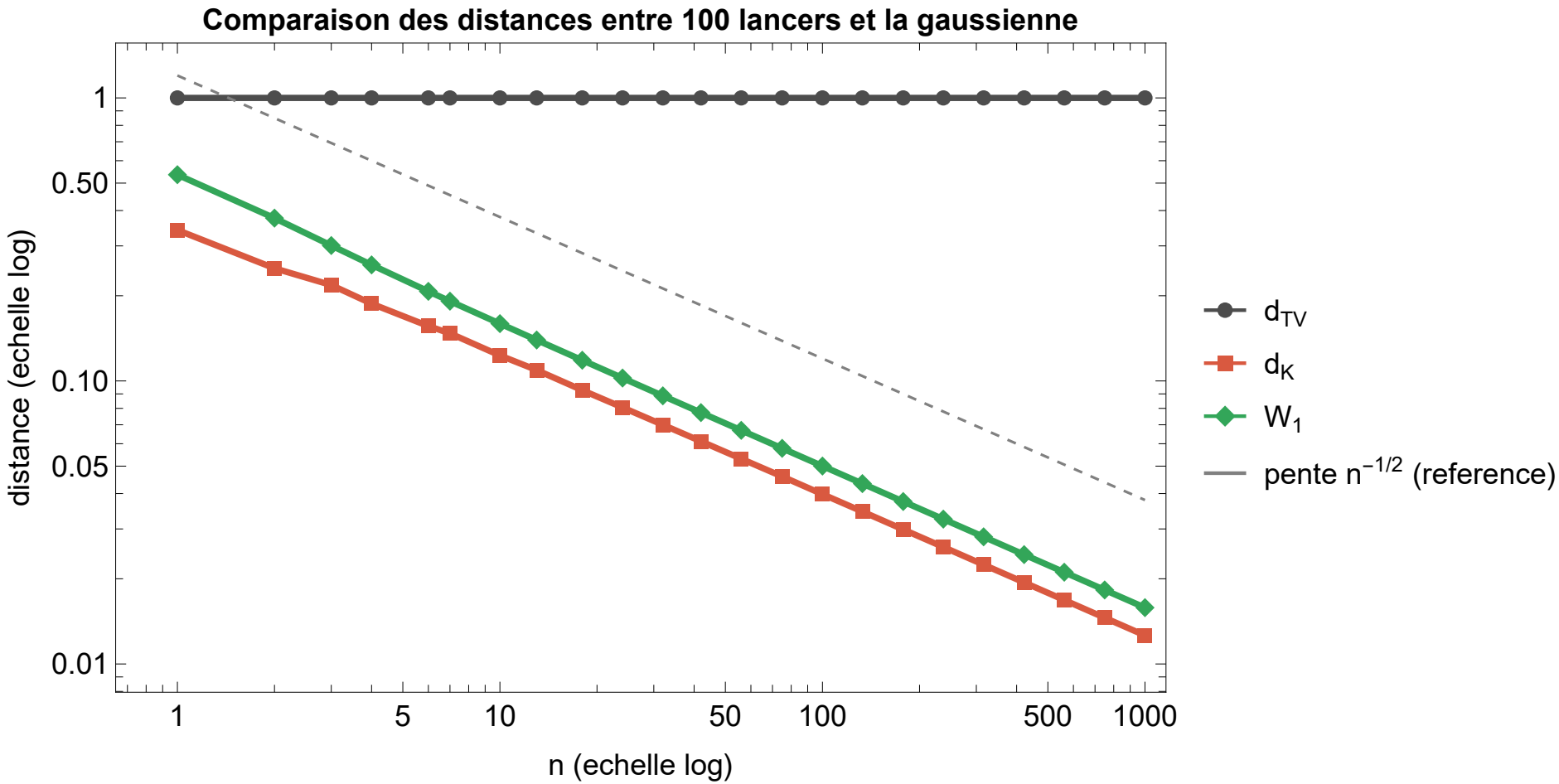


Figure: Valeur des différentes distances entre le score standardisé de l'expérience "lancer n pièce", en échelle log-log

Distances différentes, vitesses différentes - 2

La mesure de probabilité δ_x représente l'expérience où le seul outcome possible est x. Soit $(a_n)_n$ une suite convergeant vers 0. On considère δ_{a_n} et δ_0

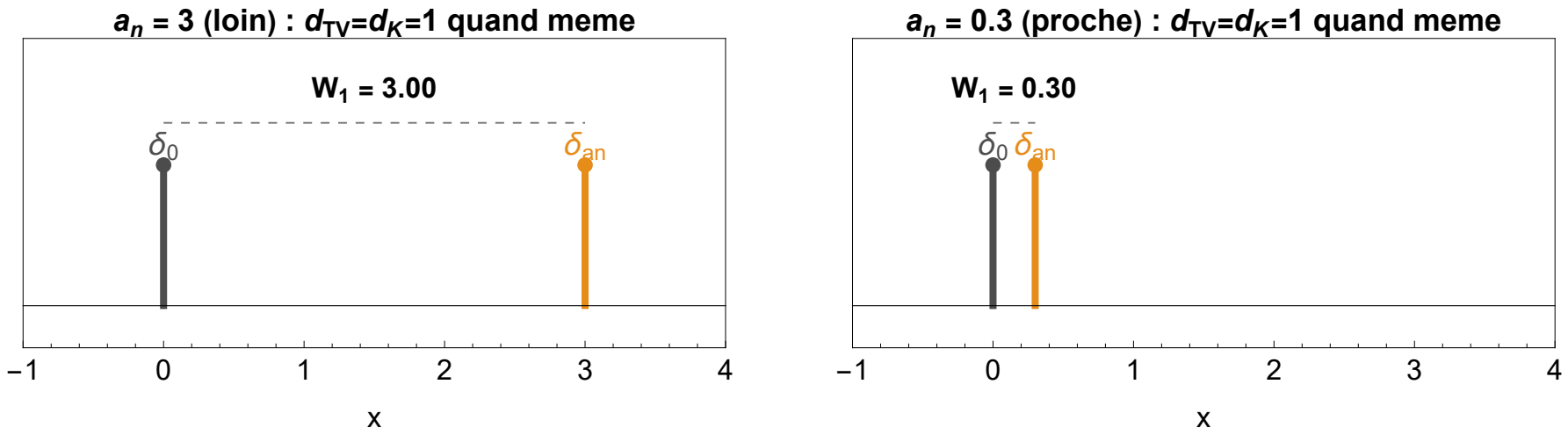


Figure: Exemple de mesure δ_x pour $x = 0$ et $x = a_n$ à deux étapes différentes de la suite a_n.

Distances différentes, vitesses différentes - 2

La mesure de probabilité δ_x représente l'expérience où le seul outcome possible est x. Soit $(a_n)_n$ une suite convergeant vers 0. On considère δ_{a_n} et δ_0

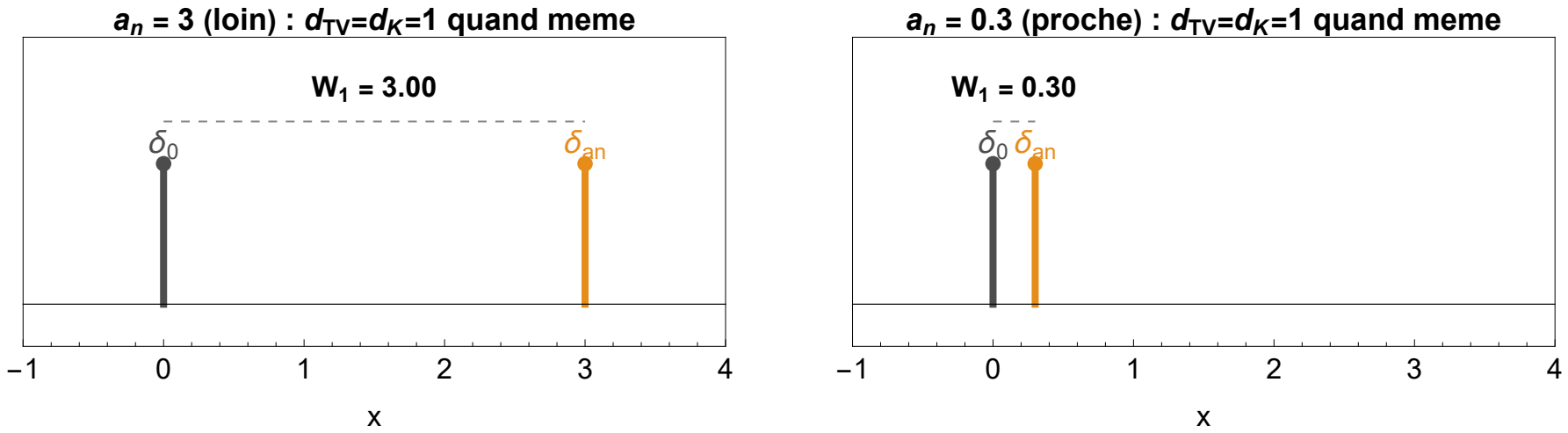


Figure: Exemple de mesure δ_x pour $x = 0$ et $x = a_n$ à deux étapes différentes de la suite a_n.

On calcule facilement

$$d_{\mathrm{W}_1}(\delta_{a_n}, \delta_0) = a_n, \quad \text{par l'interprétation de transport optimal}$$
$$d_{\mathrm{K}}(\delta_{a_n}, \delta_0) = d_{\mathrm{TV}}(\delta_{a_n}, \delta_0) = 1 \quad \text{pour } \textit{tout } n.$$

Distances différentes, vitesses différentes - 2bis

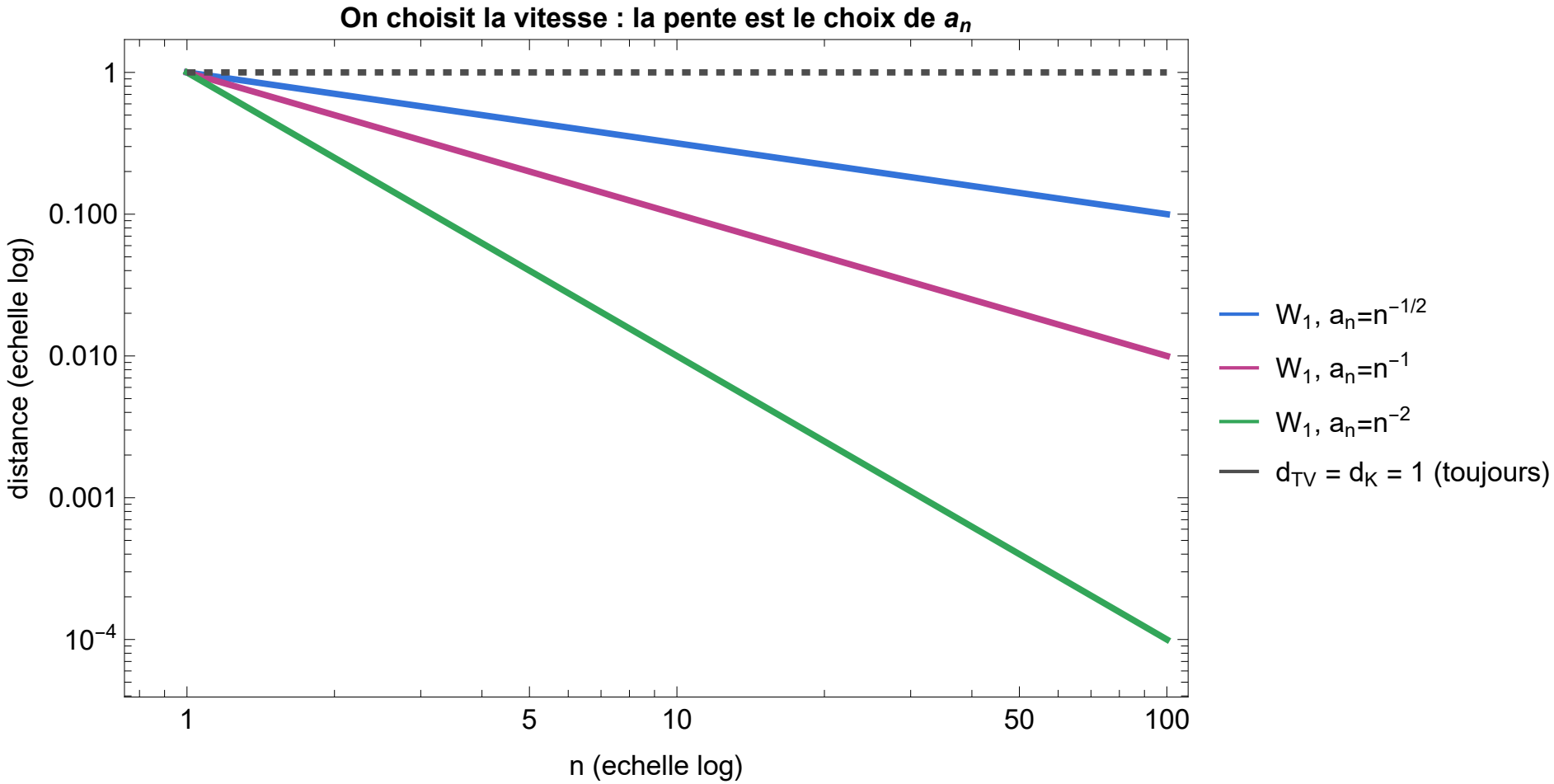


Figure: Valeur des trois distances considérées pour différentes valeurs de n et différentes suites a_n.

Distances différentes, vitesses différentes - 3

Prenons une distribution où au fil du jeu, il est de plus en plus probable d'obtenir 0, et de moins en moins probable d'obtenir une valeur c_n.

$$p_n = \left(1 - \frac{1}{n}\right) \delta_0 + \frac{1}{n} \delta_{c_n}.$$

On a facilement que $p_n \xrightarrow{\text{loi}} \delta_0$.

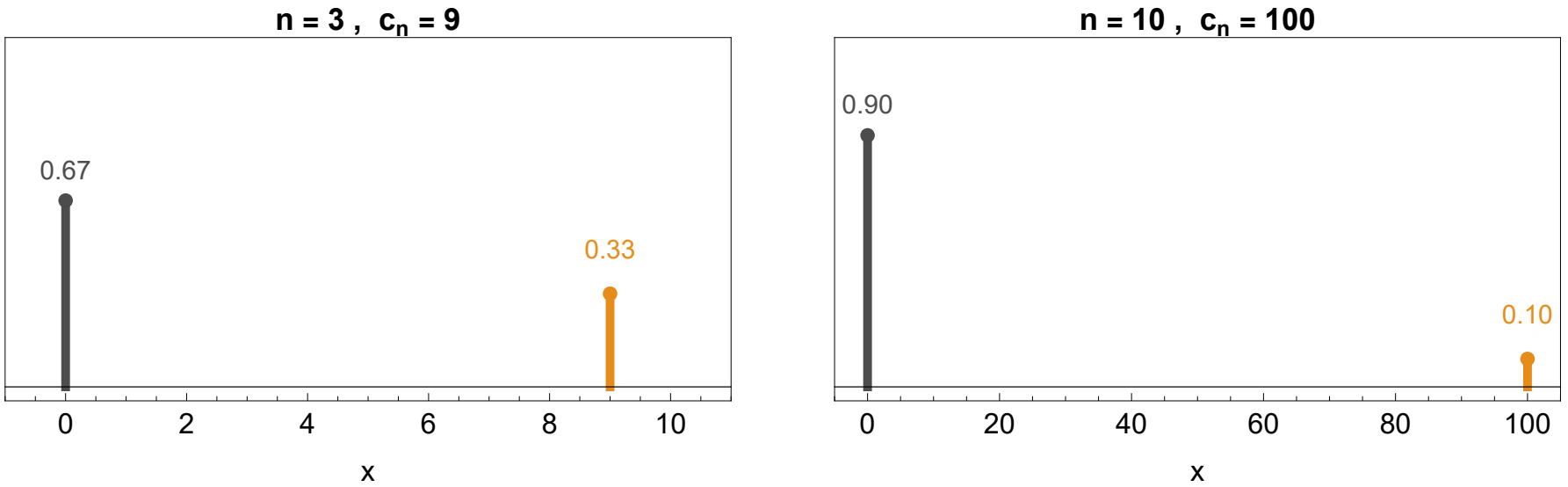


Figure: Deux exemples de la distribution p_n

Distances différentes, vitesses différentes - 3bis

On a $d_{\mathrm{TV}} = d_{\mathrm{K}} = 1/n$ ne dépendent pas de c_n. De l'autre côté, on a $d_{\mathrm{W}_1} = c_n/n$, elle, dépend entièrement du choix de c_n :

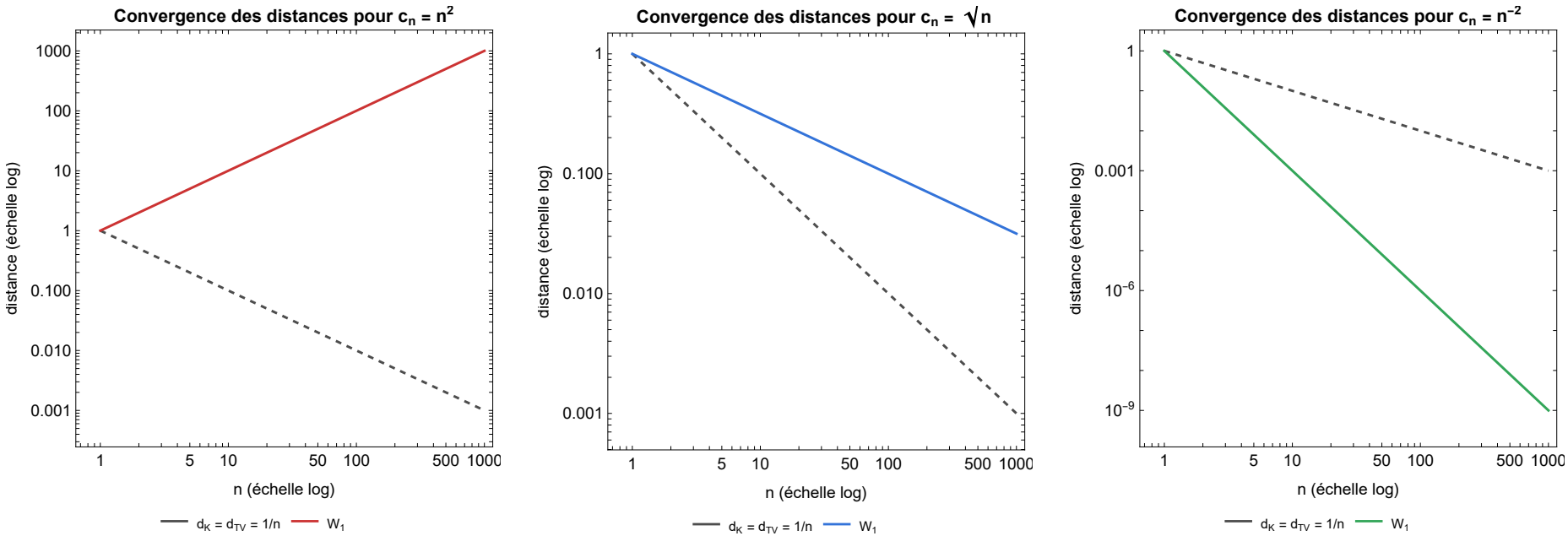


Figure: Trois régimes de convergences pour les distances considérées entre la distribution p_n et δ_0.

Outline

1. Convergences
   - Quelques jeux de probabilités
   - Ressemblances
   - Comparaisons
2. Distances entre probabilités
   - Choisir sa distance
   - Distance et vitesse
   - Encore des distances
3. Un autre aspect de la convergence
4. Conclusion

Quelle distance choisir ?

La seule vraie réponse : selon ton problème !

Quelle distance choisir ?

La seule vraie réponse : selon ton problème !

Par exemple, pour le cas d'évaluer la distance entre somme de variable indépendantes, Zolotarev a déjà réfléchi pour nous. Il souhaitait s'assurer que l'IPM ζ_s choisie a deux propriétés :

$$\zeta_s(X+Z, Y+Z) \leq \zeta_s(X,Y) \text{ et } \zeta_s(cX, cY) = |c|^s \zeta_s(X,Y),$$

et il en a construit une IPM qui respecte ces propriétés :

Définition

La distance idéale de Zolotarev d'ordre $s \in \mathbb{N} \setminus \{0\}$ est

$$\zeta_s(X,Y) = \sup_{h : \|h^{(s)}\| \leq 1} |\mathbb{E}[h(X)] - \mathbb{E}[h(Y)]|$$

La distance de Zolotarev et le TCL

Dans le cadre du TCL, et par les deux propriétés de Zolotarev :

1. $\zeta_s(X+Z, Y+Z) \leq \zeta_s(X,Y)$,
2. $\zeta_s(cX, cY) = |c|^s \zeta_s(X,Y)$,

on calcule pour Z_n une somme standardisée et N une gaussienne (comme avant),

La distance de Zolotarev et le TCL

Dans le cadre du TCL, et par les deux propriétés de Zolotarev :

1. $\zeta_s(X+Z, Y+Z) \leq \zeta_s(X,Y)$,
2. $\zeta_s(cX, cY) = |c|^s \zeta_s(X,Y)$,

on calcule pour Z_n une somme standardisée et N une gaussienne (comme avant),

$$\zeta_s(Z_n, N) \leq \sum_{i=1}^{n} \zeta_s\left(\frac{X_i}{\sqrt{n}}, \frac{N_i}{\sqrt{n}}\right)$$

La distance de Zolotarev et le TCL

Dans le cadre du TCL, et par les deux propriétés de Zolotarev :

1. $\zeta_s(X+Z, Y+Z) \leq \zeta_s(X,Y)$,
2. $\zeta_s(cX, cY) = |c|^s \zeta_s(X,Y)$,

on calcule pour Z_n une somme standardisée et N une gaussienne (comme avant),

$$\zeta_s(Z_n, N) \leq \sum_{i=1}^{n} \zeta_s\left(\frac{X_i}{\sqrt{n}}, \frac{N_i}{\sqrt{n}}\right) = n \cdot n^{-s/2} \zeta_s(X_1, N_1) \quad \text{par indépendance}$$

La distance de Zolotarev et le TCL

Dans le cadre du TCL, et par les deux propriétés de Zolotarev :

1. $\zeta_s(X+Z, Y+Z) \leq \zeta_s(X,Y)$,
2. $\zeta_s(cX, cY) = |c|^s \zeta_s(X,Y)$,

on calcule pour Z_n une somme standardisée et N une gaussienne (comme avant),

$$\zeta_s(S_n, Z) \leq \sum_{i=1}^{n} \zeta_s\left(\frac{X_i}{\sqrt{n}}, \frac{Z_i}{\sqrt{n}}\right) = n \cdot n^{-s/2} \zeta_s(X_1, Z_1) \quad \text{par indépendance}$$

La distance de Zolotarev et le TCL

Dans le cadre du TCL, et par les deux propriétés de Zolotarev :

1. $\zeta_s(X+Z, Y+Z) \leq \zeta_s(X,Y)$,
2. $\zeta_s(cX, cY) = |c|^s \zeta_s(X,Y)$,

on calcule pour Z_n une somme standardisée et N une gaussienne (comme avant),

$$\zeta_s(S_n, Z) \leq \sum_{i=1}^{n} \zeta_s\left(\frac{X_i}{\sqrt{n}}, \frac{Z_i}{\sqrt{n}}\right) = n \cdot n^{-s/2} \zeta_s(X_1, Z_1)$$

$$= n^{1-s/2} \zeta_s(X_1, Z_1)$$

La distance de Zolotarev et le TCL

Dans le cadre du TCL, et par les deux propriétés de Zolotarev :

1. $\zeta_s(X+Z, Y+Z) \leq \zeta_s(X,Y)$,
2. $\zeta_s(cX, cY) = |c|^s \zeta_s(X,Y)$,

on calcule pour Z_n une somme standardisée et N une gaussienne (comme avant),

$$\zeta_s(S_n, Z) \leq \sum_{i=1}^{n} \zeta_s\left(\frac{X_i}{\sqrt{n}}, \frac{Z_i}{\sqrt{n}}\right) = n \cdot n^{-s/2} \zeta_s(X_1, Z_1)$$
$$= n^{1-s/2} \zeta_s(X_1, Z_1)$$

Attention, ζ_s n'est finie que si les $s-1$ premiers moments de X_1 et Z_1 "matchent".

Une autre IPM

La distance en variation totale et la distance de Wasserstein ne voient pas la même chose ? Qu'à cela ne tienne, regardons les ensemble !

Une autre IPM

La distance en variation totale et la distance de Wasserstein ne voient pas la même chose ? Qu'à cela ne tienne, regardons les ensemble !

Definition

La distance de Fortet-Mourier d_{FM} est l'IPM associée à la classe

$$\mathcal{H} = \{h : \|h\| \leq 1 \text{ et } \|h'\| \leq 1\}$$

Elle a la propriété de "métriser" la convergence en loi, c-à-d :

$$X_n \xrightarrow{\text{loi}} Y \Leftrightarrow d_{\mathrm{FM}}(X_n, Y) \to 0.$$

Une autre IPM

La distance en variation totale et la distance de Wasserstein ne voient pas la même chose ? Qu'à cela ne tienne, regardons les ensemble !

Definition

La distance de Fortet-Mourier d_{FM} est l'IPM associée à la classe

$$\mathcal{H} = \{h : \|h\| \leq 1 \ et \ \|h'\| \leq 1\}$$

Elle a la propriété de "métriser" la convergence en loi, c-à-d :

$$X_n \xrightarrow{\text{loi}} Y \Leftrightarrow d_{\mathrm{FM}}(X_n, Y) \to 0.$$

Un autre aspect de la convergence

Les constantes

Théorème (Berry–Esseen)

Soit Z_n, la standardisation de la somme des X_i, et N une gaussienne. Supposons que ρ (la moyenne de $|X_1 - \mu|^3$) soit finie. Il existe alors une constante universelle C – qui ne dépend ni de n, ni de la loi de X_1 – telle que

$$d_{\mathrm{K}}(Z_n, N) \ \leq \ \frac{C\,\rho}{\sigma^3\sqrt{n}}.$$

Les constantes

Théorème (Berry–Esseen)

Soit Z_n, la standardisation de la somme des X_i, et N une gaussienne. Supposons que ρ (la moyenne de $|X_1 - \mu|^3$)soit finie. Il existe alors une constante universelle C – qui ne dépend ni de n, ni de la loi de X_1 – telle que

$$d_{\mathrm{K}}(Z_n, N) \leq \frac{C\,\rho}{\sigma^3\sqrt{n}}.$$

Les constantes

Théorème (Berry–Esseen)

Soit Z_n, la standardisation de la somme des X_i, et N une gaussienne. Supposons que ρ (la moyenne de $|X_1 - \mu|^3$)soit finie. Il existe alors une constante universelle C – qui ne dépend ni de n, ni de la loi de X_1 – telle que

$$d_{\mathrm{K}}(Z_n, N) \leq \frac{C\,\rho}{\sigma^3\sqrt{n}}.$$

Les constantes

Théorème (Berry–Esseen)

Soit Z_n, la standardisation de la somme des X_i, et N une gaussienne. Supposons que ρ (la moyenne de $|X_1 - \mu|^3$)soit finie. Il existe alors une constante universelle C – qui ne dépend ni de n, ni de la loi de X_1 – telle que

$$d_{\mathrm{K}}(Z_n, N) \leq \frac{C\,\rho}{\sigma^3\sqrt{n}}.$$

Et à quoi ça sert ?

Imaginons qu'on ait besoin de s'assurer un certain degré de précision en distance de Kolmogorov (disons 1%), quelle est le nombre de "lancer de pièce" nécessaire pour que l'approximation soit du bon ordre de précision ?

Et à quoi ça sert ?

Imaginons qu'on ait besoin de s'assurer un certain degré de précision en distance de Kolmogorov (disons 1%), quelle est le nombre de "lancer de pièce" nécessaire pour que l'approximation soit du bon ordre de précision ?

Dans ce cadre, $\sigma = 0{,}5$ et $\rho = 0{,}125$, donc $\rho/\sigma^3 = 1$ et Berry–Esseen se simplifie en

$$d_K(Z_n, \mathcal{N}) \;\leq\; \frac{C}{\sqrt{n}}.$$

Pour garantir $d_K \leq 1\%$, il faut $n_0 = (100\,C)^2$ lancers.

Constante C	Origine	n_0 pour $d_K \leq 1\%$
$7{,}59$	Esseen, preuve originale (1942)	$\approx 576\,000$
$0{,}7975$	van Beek (1972)	$\approx 6\,360$
$0{,}7655$	Shiganov (1986)	$\approx 5\,860$
$0{,}4748$	Shevtsova (2014), meilleure borne connue	$\approx 2\,250$

Comment évaluer cette constante ?

La méthode de Stein se base sur les idées suivantes :

Comment évaluer cette constante ?

La méthode de Stein se base sur les idées suivantes :

- Une égalité d'espérances caractéristique

$$N \sim \mathcal{N}(0,1) \quad \Leftrightarrow \quad \mathbb{E}[Nf(N)] = \mathbb{E}[f'(N)], \tag{1}$$

Comment évaluer cette constante ?

La méthode de Stein se base sur les idées suivantes :

- Une égalité d'espérances caractéristique

$$N \sim \mathcal{N}(0,1) \quad \Leftrightarrow \quad \mathbb{E}[Nf(N)] = \mathbb{E}[f'(N)], \tag{1}$$

- D'où on tire une *Stein equation*,

$$f'(x) - xf(x) = h(x) - \mathbb{E}[h(N)]. \tag{2}$$

Comment évaluer cette constante ?

La méthode de Stein se base sur les idées suivantes :

- Une égalité d'espérances caractéristique

$$N \sim \mathcal{N}(0,1) \quad \Leftrightarrow \quad \mathbb{E}[Nf(N)] = \mathbb{E}[f'(N)], \tag{1}$$

- D'où on tire une *Stein equation*,

$$f'(x) - xf(x) = h(x) - \mathbb{E}[h(N)]. \tag{2}$$

- Un principe de transfert,

$$d_{\mathcal{H}}(X,N) = \sup_{h\in\mathcal{H}} |\,\mathbb{E}[h(X)] - \mathbb{E}[h(N)]| = \sup_{h\in\mathcal{H}} |\,\mathbb{E}[f_h'(X) - Xf_h(X)]|.$$

Comment évaluer cette constante ?

La méthode de Stein se base sur les idées suivantes :

- Une égalité d'espérances caractéristique

$$N \sim \mathcal{N}(0,1) \quad \Leftrightarrow \quad \mathbb{E}[Nf(N)] = \mathbb{E}[f'(N)], \tag{1}$$

- D'où on tire une *Stein equation*,

$$f'(x) - xf(x) = h(x) - \mathbb{E}[h(N)]. \tag{2}$$

- Un principe de transfert,

$$d_{\mathcal{H}}(X,N) = \sup_{h\in\mathcal{H}} |\,\mathbb{E}[h(X)] - \mathbb{E}[h(N)]| = \sup_{h\in\mathcal{H}} |\,\mathbb{E}[f_h'(X) - Xf_h(X)]|.$$

- De bonnes propriétés de f_h la solution à (2) et une utilisation intelligente des propriétés du design experimental.

Et alors ?

Et alors ?

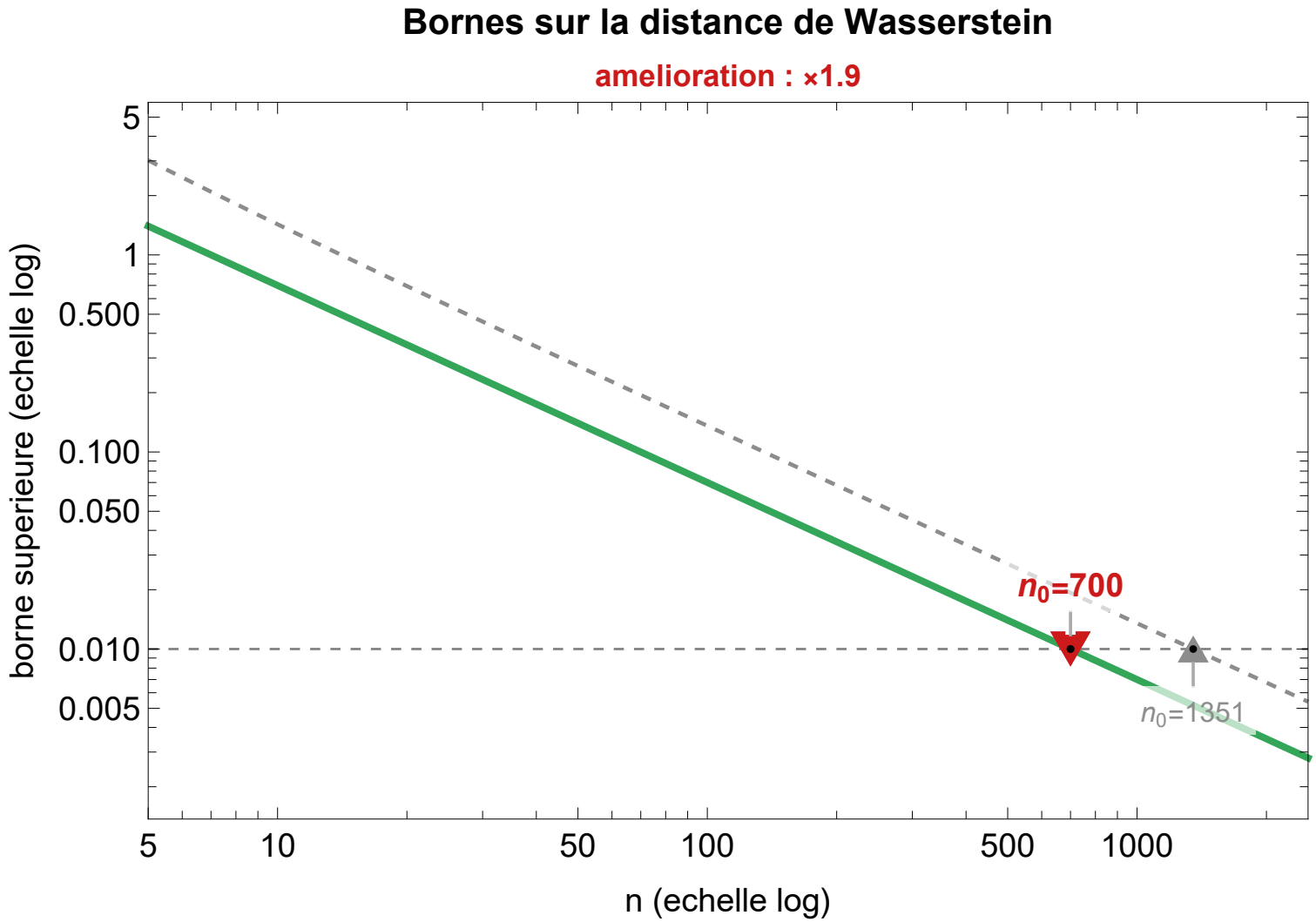
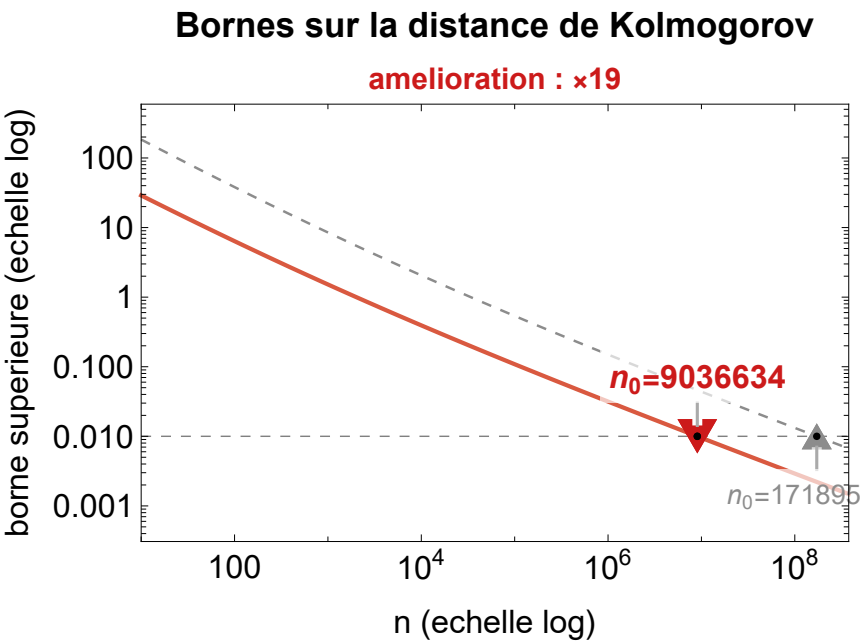


Figure: Distance de Wasserstein entre le temps de dernier retour à 0 d'une marche aléatoire et son approximation par une loi Beta selon [**GR13**] (pointillé) et selon [**RS26**] (vert)



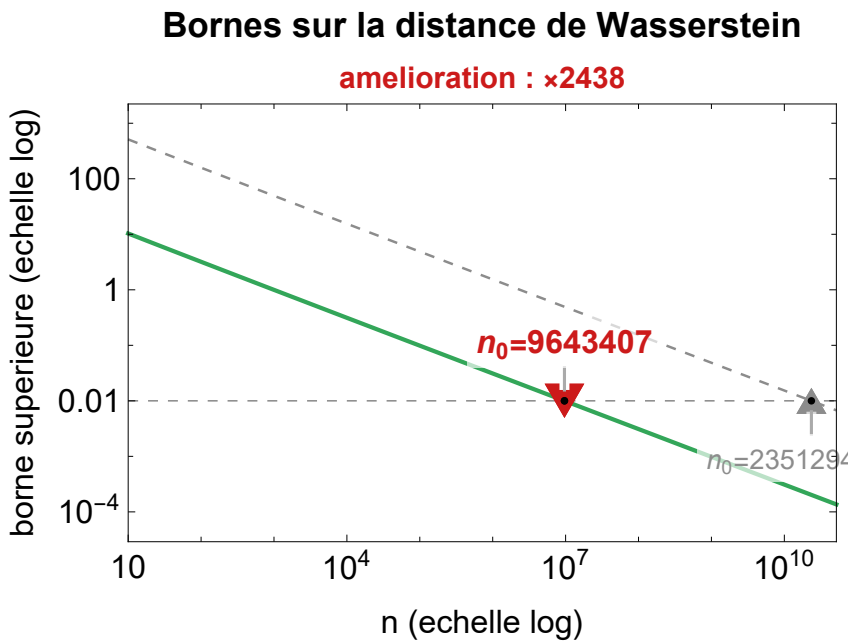


Figure: Distance de Kolmogorov (gauche) et de Wasserstein (droite) entre le spin dans le modèle de Curie-Weiss à température critique d'après [**CS11**] (pointillé) et selon [**RS26**] (rouge, vert)

Conclusion

Conclusion

1. Une distance, c'est une question qu'on pose à une loi. C'est notre outil de mesure !

Conclusion

1. Une distance, c'est une question qu'on pose à une loi. C'est notre outil de mesure !
2. De la question dépend la vitesse de rapprochement

Conclusion

1. Une distance, c'est une question qu'on pose à une loi. C'est notre outil de mesure !
2. De la question dépend la vitesse de rapprochement
3. Choisir n'est pas gratuit (ζ_s demande des guaranties sur la ressemblance des deux lois)

Conclusion

1. Une distance, c'est une question qu'on pose à une loi. C'est notre outil de mesure !
2. De la question dépend la vitesse de rapprochement
3. Choisir n'est pas gratuit (ζ_s demande des guaranties sur la ressemblance des deux lois)
4. Sans une bonne constante, un bon taux de convergence n'est pas si utile

Merci pour votre attention !