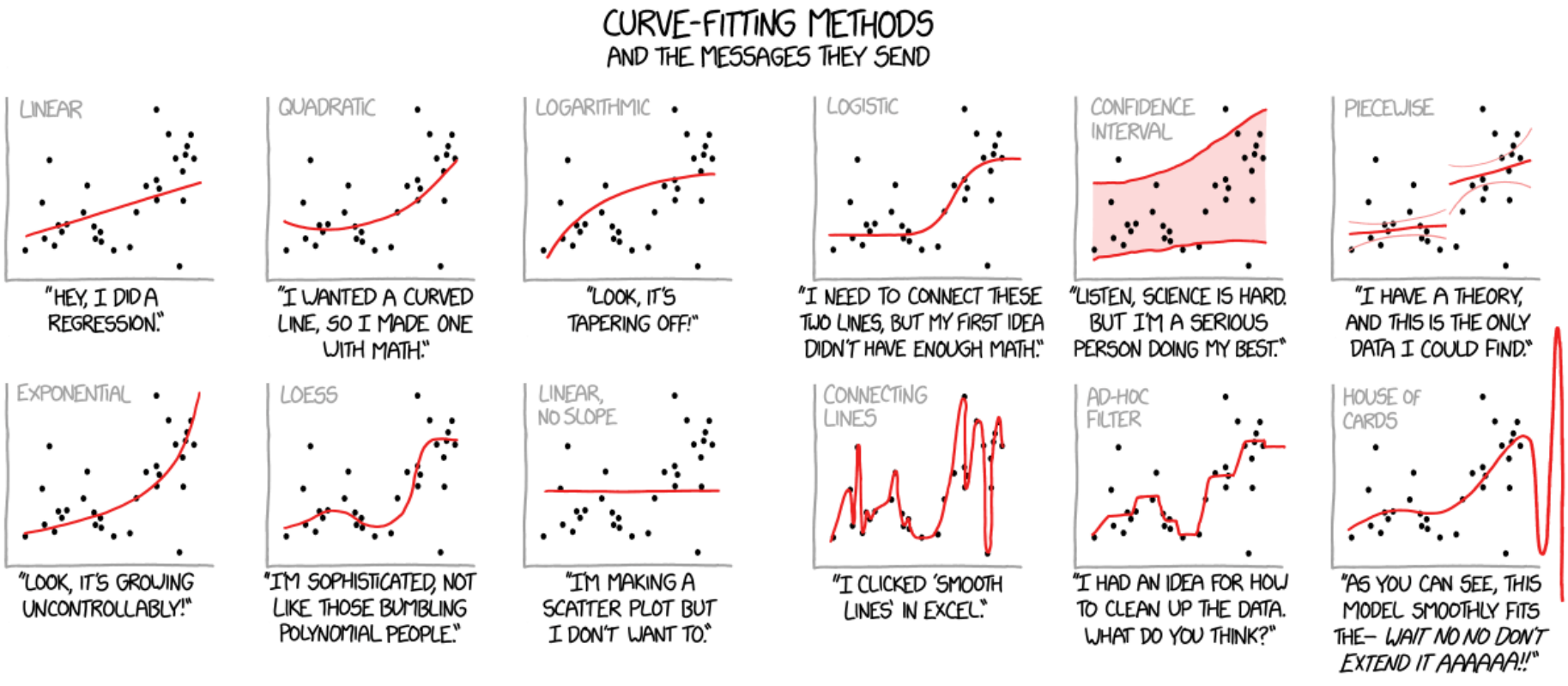
Statistique

Benjamin Bobbia

ISAE

Statistique inférentielle



Curve-Fitting, XKCD, xkcd.com/2048

0.0 Notions élémentaires

Vocabulaire

Dans le cadre de la statistique inférentielle, nous considérerons les observations $x_1, \dots, x_n \in \mathcal{E}$ du phénomène étudié comme des **réalisations** de variables aléatoires $X_1, \dots, X_n$ à valeurs dans $\mathcal{E}$. Ces données sont appelées un **échantillon**.

Bien que les variables aléatoires $X_1, \dots, X_n$ seront souvent supposées *i.i.d.* dans la suite, ce n'est pas toujours le cas en statistique inférentielle. De telles hypothèses forment le **cadre statistique** du problème considéré.

Toute quantité calculée **uniquement à partir de l'échantillon** est appelée un **estimateur**. Il s'agit donc d'une **fonction des observations** et cela s'exprime comme une réalisation d'une fonction des variables $X_1, \dots, X_n$.

Il est **crucial** de comprendre que les mesures effectuées sur l'échantillon sont des **réalisations** de variables aléatoires sous-jacentes.

Moyenne empirique

Dans le cas de variables aléatoires réelles $X_1, \ldots, X_n$, l'échantillon est constitué de valeurs observées $x_1, \ldots, x_n \in \mathbb{R}$.

La valeur moyenne $\overline{x}_n$ est donc une **réalisation de la variable aléatoire** $\overline{X}_n$,

$$\overline{X}_n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

La variable aléatoire $\overline{X}_n$ est appelée la **moyenne empirique**.

La moyenne empirique est un **estimateur**.

Remarque : « être un estimateur » ne signifie par « être un estimateur *de quelque chose* ». Cela signifie uniquement « être construit à partir des observations ».

Moyenne empirique

Si les variables $X_1, \ldots, X_n$ ont la **même loi** et qu'elle admet une espérance $\mathbb{E}[X_1] = m \in \mathbb{R}$, alors

$$\mathbb{E}[\overline{X}_n] = \frac{1}{n} \sum_{k=1}^{n} \mathbb{E}[X_k] = \frac{nm}{n} = m.$$

Biais

Le **biais** d'un estimateur T_n pour estimer un **paramètre** $t \in \mathbb{R}$ est l'écart entre l'espérance de T_n et sa cible,

$$b(T_n) = \mathbb{E}[T_n] - t.$$

Si $b(T_n) = 0$, l'estimateur est dit **sans biais** pour estimer t.
Si $b(T_n) \to 0$ quand la taille n de l'échantillon tend vers l'infini, l'estimateur est dit **asymptotiquement sans biais** pour estimer t.

La moyenne empirique est **sans biais** pour estimer la moyenne m.

Moyenne empirique

Si les variables $X_1, \ldots, X_n$ sont *i.i.d.* et admettent une variance commune $\mathrm{Var}(X_1) = \sigma^2$, alors, par indépendance,

$$\mathrm{Var}(\overline{X}_n) = \mathbb{E}\left[\left(\overline{X}_n - \mathbb{E}[\overline{X}_n]\right)^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - m)\right)^2\right]$$

$$= \frac{1}{n^2}\sum_{k=1}^{n}\sum_{k'=1}^{n}\mathbb{E}\left[(X_k - m)(X_{k'} - m)\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Convergence en moyenne quadratique

Un estimateur T_n **converge en moyenne quadratique** vers un paramètre $t \in \mathbb{R}$ si l'espérance de l'écart au carré entre T_n et sa cible tend vers 0 quand la taille n de l'échantillon tend vers l'infini,

$$\mathbb{E}\left[(T_n - t)^2\right] = b(T_n)^2 + \mathrm{Var}(T_n) \xrightarrow[n\to\infty]{} 0.$$

La moyenne empirique **converge en moyenne quadratique** vers m.

Compromis biais-variance

Si les variables $X_1, \ldots, X_n$ sint i.i.d de variance finie commune σ^2, on a pour tout estimateur T_n de variance finie

$$\mathbb{E}\left[(T_n - t)^2\right] = b(T_n)^2 + \mathrm{Var}(T_n).$$

Pour avoir un "bon" estimateur il faut donc

- Un biais faible, i.e précision
- Une variance faible, i.e faible variabilité

Compromis biais-variance

Si les variables $X_1, \ldots, X_n$ sint i.i.d de variance finie commune σ^2, on a pour tout estimateur T_n de variance finie

$$\mathbb{E}\left[(T_n - t)^2\right] = b(T_n)^2 + \mathrm{Var}(T_n).$$

Pour avoir un "bon" estimateur il faut donc

- Un biais faible, i.e précision
- Une variance faible, i.e faible variabilité

Problème

En général faire décroitre le biais entraine une augmentation de la variance.

Les méthodes permettant de choisir peuvent être de la **régularisation**, **validation croisée**, ..., mais vous en reparlerez plus tard.

Borne de Cramer-Rao

La borne de Fréchat-Darmois-Cramer-Rao, met en lumière qu'**il n'existe pas** d'estimateur "parfait" c'est à dire tel que $b(T_n) = 0$ et $Var(T_n) = 0$. Pour T_n un estimateur d'un paramètre $\theta \in \mathbb{R}$ construit sur un échantillon $X_1, \dots, X_n$ i.i.d on défini

Information de Fisher

$$I(\theta) = -\mathbb{E}_\theta \left[\partial_\theta^2 \ln(f_\theta(X_1)) \right]$$

où f_θ désigne la densité de X_1.

Borne de Cramer-Rao

Si T_n est un estimateur **sans biais** de θ alors

$$\mathrm{Var}_\theta(\hat{\theta}_n) \geqslant \frac{1}{nI(\theta)}.$$

Variance empirique

Dans le cas de variables aléatoires réelles $X_1, ;X_n$, i.i.d de variance finie commune σ^2, on peut estimer Σ^2 par

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2.$$

On l'appel la **variance empirique**

Biais de la variance empirique

On a

$$\mathbb{E}(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2$$

Variance empirique

Dans le cas de variables aléatoires réelles $X_1, ;X_n$, i.i.d de variance finie commune σ^2, on peut estimer Σ^2 par

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2.$$

On l'appel la **variance empirique**

Biais de la variance empirique

On a

$$\mathbb{E}(\hat{\sigma}_n^2) = \frac{n-1}{n}\sigma^2$$

- L'estimateur $\hat{\sigma}_n^2$ est asymptotiquement sans biais
- **MAIS** il est biaisé.

Variance empirique

Dans la pratique on utilisera plutôt

$$\hat{s}_n^2 = \frac{n}{n-1}\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2.$$

Attention

- Cette estimateur est sans biais donc (souvent) préférable dans la pratique.
- C'est cet estimateur qui est implémenté dans la plupart des logiciels (R, Python, Statistica, SAS,...)
- Le facteur $\frac{n-1}{n}$ est peu impactant lorsque n est grand mais a son importance pour des petit échantillon.

Inégalité de Bienaymé-Tchebychev

Si T est une variable aléatoire qui admet une espérance et une variance finies, alors

$$\forall \varepsilon > 0,\ \mathbb{P}\left(|T - \mathbb{E}[T]| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}(T)}{\varepsilon^2}.$$

Preuve : il suffit d'introduire la variable aléatoire binaire B définie par

$$B = \begin{cases} 1 & \text{si } |T - \mathbb{E}[T]| \geqslant \varepsilon, \\ 0 & \text{sinon.} \end{cases}$$

La variance de T se décompose alors comme suit,

$$\begin{aligned} \mathrm{Var}(T) &= \mathbb{E}[(T - \mathbb{E}[T])^2] \\ &= \mathbb{E}[(T - \mathbb{E}[T])^2 B] + \mathbb{E}[(T - \mathbb{E}[T])^2 (1 - B)] \\ &\geqslant \mathbb{E}[(T - \mathbb{E}[T])^2 B] \\ &\geqslant \mathbb{E}[\varepsilon^2 B] = \varepsilon^2 \mathbb{P}(|T - \mathbb{E}[T]| \geqslant \varepsilon) \end{aligned}$$

car B suit la loi de Bernoulli de paramètre $p = \mathbb{P}(|T - \mathbb{E}[T]| \geqslant \varepsilon)$.

Inégalité de Bienaymé-Tchebychev

Si T est une variable aléatoire qui admet une espérance et une variance finies, alors

$$\forall \varepsilon > 0,\ \mathbb{P}\left(|T - \mathbb{E}[T]| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}(T)}{\varepsilon^2}.$$

L'intérêt de ce type d'inégalité est de **quantifier la variabilité** de T autour de son espérance. En effet, en posant $\alpha = \mathrm{Var}(T)/\varepsilon^2$, nous obtenons

$$\mathbb{P}\left(|T - \mathbb{E}[T]| \geqslant \frac{\sqrt{\mathrm{Var}(T)}}{\sqrt{\alpha}}\right) \leqslant \alpha.$$

Autrement dit, si $\alpha \in]0, 1[$, nous en déduisons que

$$\mathbb{E}[T] \in \left]T - \sqrt{\frac{\mathrm{Var}(T)}{\alpha}}; T + \sqrt{\frac{\mathrm{Var}(T)}{\alpha}}\right[$$

avec une probabilité supérieure à $1 - \alpha$.

Inégalité de Bienaymé-Tchebychev

Si T est une variable aléatoire qui admet une espérance et une variance finies, alors

$$\forall \varepsilon > 0,\ \mathbb{P}\left(|T - \mathbb{E}[T]| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}(T)}{\varepsilon^2}.$$

Dans le cas de variables réelles $X_1, \ldots, X_n$ *i.i.d.* avec $\mathbb{E}[X_1] = m$ et $\mathrm{Var}(X_1) = \sigma^2$, cette inégalité appliquée à $T = \overline{X}_n$ donne, pour tout $\alpha \in]0,1[$,

$$m \in \left]\overline{X}_n - \sqrt{\frac{\sigma^2}{\alpha n}}; \overline{X}_n + \sqrt{\frac{\sigma^2}{\alpha n}}\right[$$

avec probabilité supérieure à $1 - \alpha$.

Inégalité de Bienaymé-Tchebychev

Si T est une variable aléatoire qui admet une espérance et une variance finies, alors

$$\forall \varepsilon > 0,\ \mathbb{P}\left(|T - \mathbb{E}[T]| \geqslant \varepsilon\right) \leqslant \frac{\mathrm{Var}(T)}{\varepsilon^2}.$$

Intervalle de confiance

Soient A_n et B_n des **estimateurs** réels avec $A_n < B_n$ presque sûrement. Pour un paramètre $t \in \mathbb{R}$, si il existe $\alpha \in]0,1[$ tel que

$$\mathbb{P}\left(t \in]A_n; B_n[\right) \geqslant 1 - \alpha$$

alors $]A_n, B_n[$ est appelé **intervalle de confiance** de niveau $1 - \alpha$ pour le paramètre t.

Il s'agit d'un intervalle dont les bornes sont aléatoires.

Exemple simulé : estimation d'une proportion

Une généticienne étudie une mutation présente seulement dans le génome d'une partie de la population. Afin de mesurer la proportion $p \in]0,1[$ **inconnue** de la population qui présente cette mutation, elle prélève **uniformément** au hasard n individu **avec remise** et note le résultat,

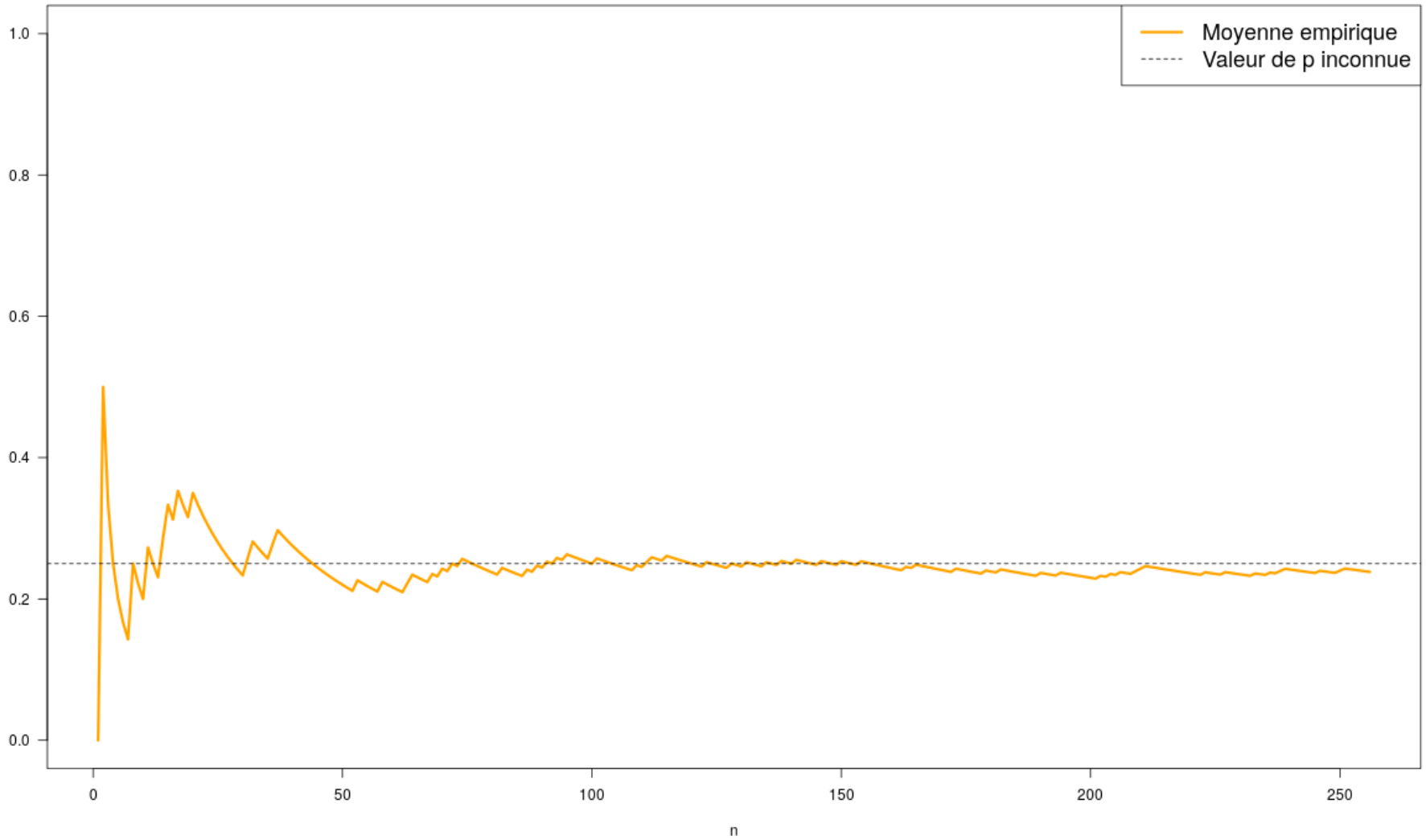
$$\forall k \in \{1,\dots,n\},\ X_k = \begin{cases} 1 & \text{si le } k\text{ème individu présente la mutation,} \\ 0 & \text{sinon.} \end{cases}$$

Les variables $X_1,\dots,X_n$ sont *i.i.d.* de loi de Bernoulli $\mathcal{B}(p)$.

Puisque $\mathbb{E}[X_1] = p$, la moyenne empirique $\overline{X}_n$ peut être utilisée comme estimateur sans biais de p.

Les données de cet exemple sont simulées avec $p = 0.25$.

Exemple simulé : estimation d'une proportion



Exemple simulé : estimation d'une proportion

Pour établir la fourchette de l'estimation, elle considère l'intervalle de confiance de niveau $1-\alpha \in]0,1[$ donné par

$$\left]\overline{X}_n - \sqrt{\frac{p(1-p)}{\alpha n}}; \overline{X}_n + \sqrt{\frac{p(1-p)}{\alpha n}}\right[.$$

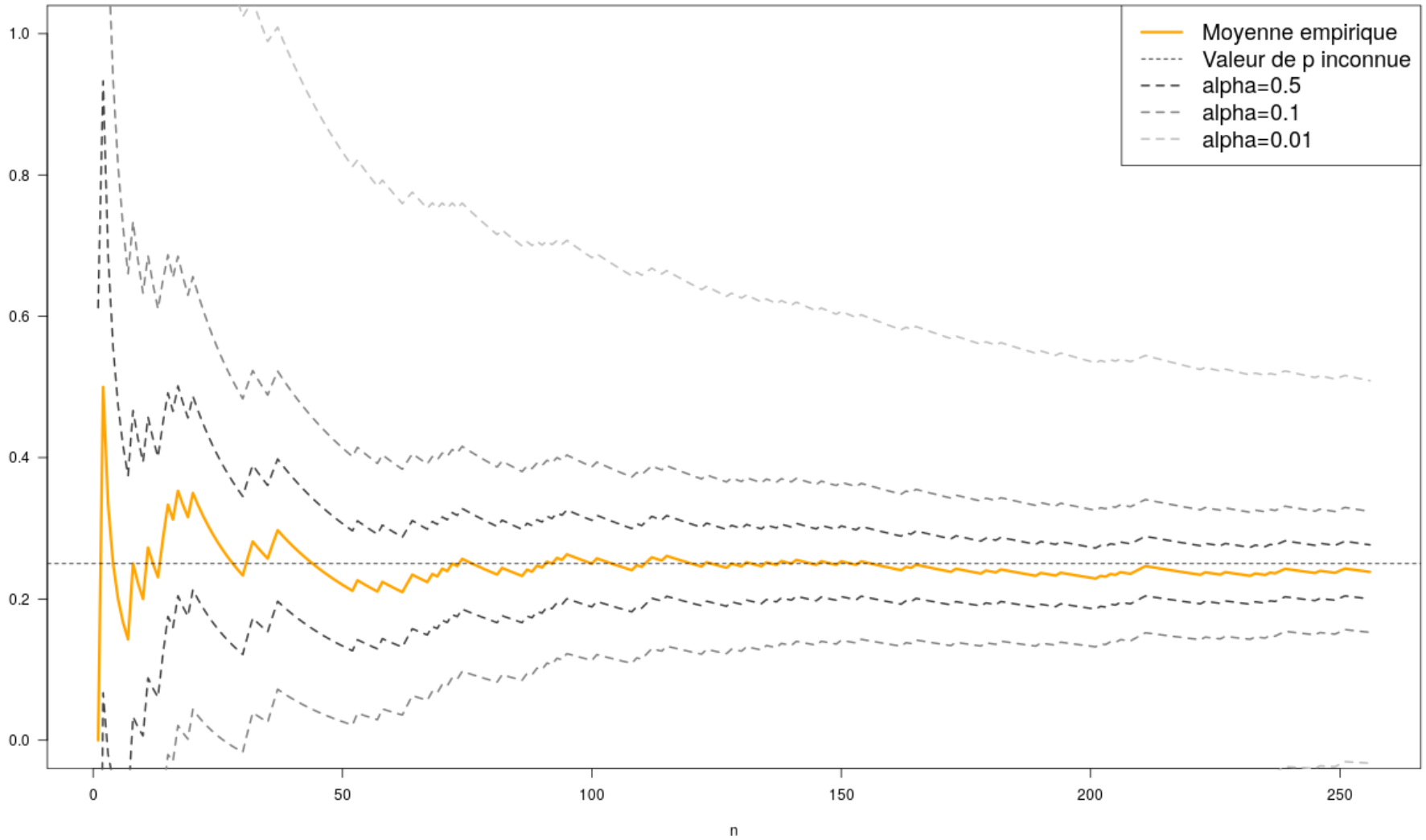
car $\mathrm{Var}(X_1) = p(1-p)$.

Dans cet exemple, $p(1-p) = 0.1875$. Par exemple, pour $\alpha = 0.1$, nous obtenons que

$$p \in \left]\overline{X}_n - \sqrt{\frac{1.875}{n}}; \overline{X}_n + \sqrt{\frac{1.875}{n}}\right[$$

avec une probabilité supérieure à $1-\alpha = 90\%$.

Exemple simulé : estimation d'une proportion



Exemple simulé : estimation d'une proportion

STOP ! ! !

Les intervalles précédents donnés par

$$\left]\overline{X}_n - \sqrt{\frac{p(1-p)}{\alpha n}}; \overline{X}_n + \sqrt{\frac{p(1-p)}{\alpha n}}\right[$$

ne sont **pas des intervalles de confiance**. En effet, les bornes dépendent du paramètre p inconnu et pas uniquement des observations. Ce ne sont pas des estimateurs.

Exemple simulé : estimation d'une proportion

Il existe plusieurs façons de contourner ce problème :

1. majorer la variance (si possible),

Dans le cas d'une proportion, nous savons que

$$\forall p \in]0,1[,\ p(1-p) \leqslant \frac{1}{4}.$$

Nous obtenons un intervalle de confiance de niveau $1-\alpha \in]0,1[$ défini par

$$IC_1 = \left]\overline{X}_n - \frac{1}{2\sqrt{\alpha n}}; \overline{X}_n + \frac{1}{2\sqrt{\alpha n}}\right[$$

Exemple simulé : estimation d'une proportion

Il existe plusieurs façons de contourner ce problème :

1. majorer la variance (si possible),
2. résoudre explicitement la dépendance (si possible),

Pour une proportion, cela se ramène à une inéquation du second degré,

$$|\overline{X}_n - p| < \sqrt{\frac{p(1-p)}{\alpha n}} \iff (\overline{X}_n - p)^2 < \frac{p(1-p)}{\alpha n}$$
$$\iff (1+\alpha n)p^2 - (2\alpha n\overline{X}_n + 1)p + \alpha n\overline{X}_n^2 < 0.$$

Nous obtenons un intervalle de confiance de niveau $1-\alpha \in]0,1[$ défini par

$$IC_2 = \left] \frac{2\alpha n\overline{X}_n + 1 - \sqrt{\Delta}}{2(1+\alpha n)}; \frac{2\alpha n\overline{X}_n + 1 + \sqrt{\Delta}}{2(1+\alpha n)} \right[$$

avec $\Delta = 1 + 4\alpha n\overline{X}_n(1-\overline{X}_n)$.

Exemple simulé : estimation d'une proportion

Il existe plusieurs façons de contourner ce problème :

1. majorer la variance (si possible),
2. résoudre explicitement la dépendance (si possible),
3. estimer la variance.

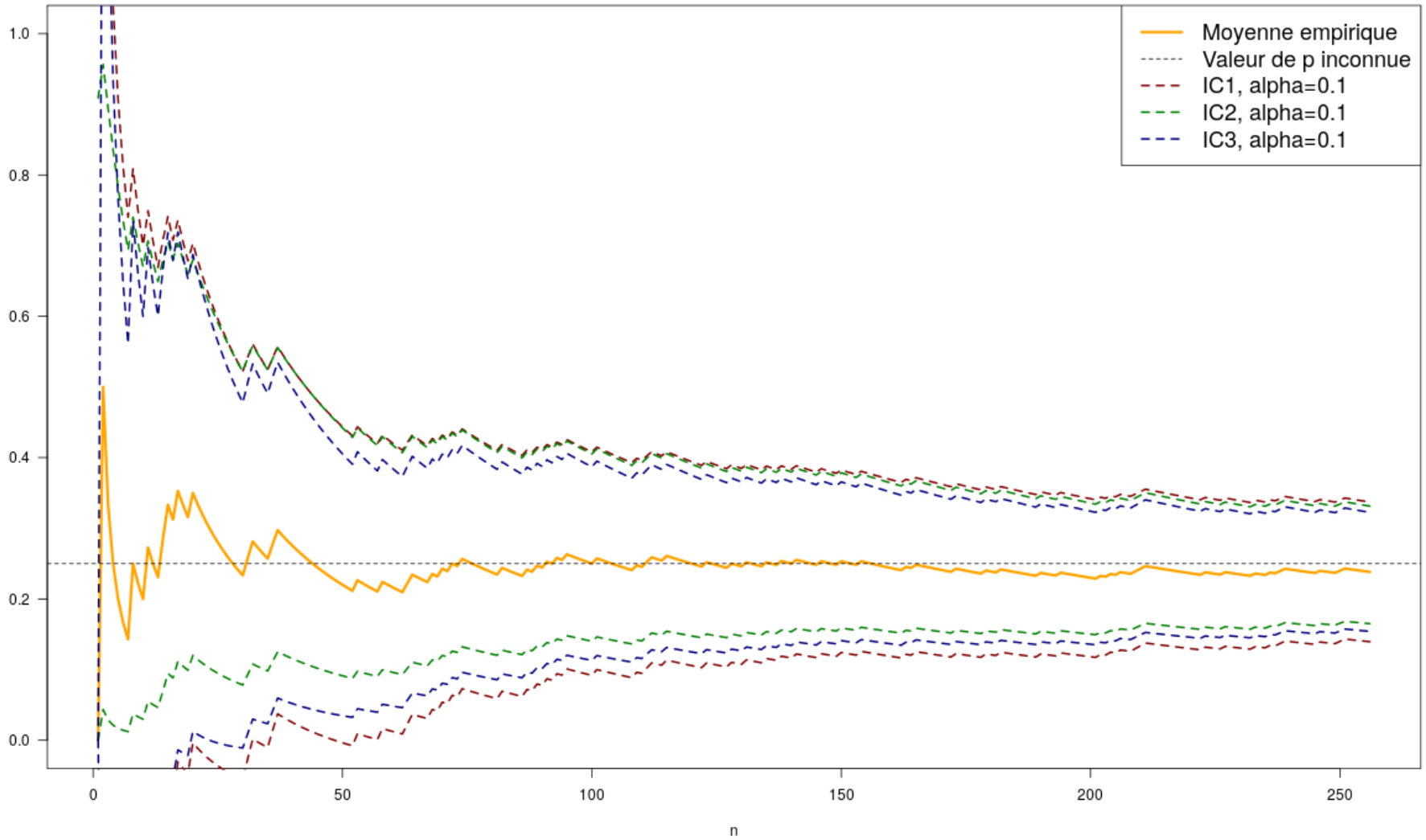
Si nous disposons d'un **estimateur de la variance** $\hat{\sigma}_n^2$, il est possible de l'utiliser pour définir l'intervalle

$$IC_3 = \left]\overline{X}_n - \sqrt{\frac{\hat{\sigma}_n^2}{\alpha n}}; \overline{X}_n + \sqrt{\frac{\hat{\sigma}_n^2}{\alpha n}}\right[.$$

Cette méthode est largement utilisée en pratique mais, a priori, elle **ne garantit plus le niveau** $1-\alpha$. Pour une proportion, nous pouvons prendre

$$\hat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 \qquad \text{ou} \qquad \hat{\sigma}_n^2 = \overline{X}_n(1-\overline{X}_n).$$

Exemple simulé : estimation d'une proportion



Convergence de variables aléatoires

Soit $(X_n)_{n\in\mathbb{N}}$ et X des variable aléatoires défini sur un espace probabilisé E. On dit que

- On dit que X_n converge **presque sûrement** vers X si

$$\mathbb{P}(X_n \underset{n\to\infty}{\longrightarrow} X) = 1$$

  On le note $X_n \underset{n\to\infty}{\overset{p.s}{\longrightarrow}} X$.
- On dit que X_n converge **en probabilité** vers X si pour tout $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}(\|X_n - X\| > \varepsilon) = 0.$$

  On le note $X_n \underset{n\to\infty}{\overset{\mathbb{P}}{\longrightarrow}} X$.

Convergence de variables aléatoires

Soit $(X_n)_{n\in\mathbb{N}}$ et X des variable aléatoires défini sur un espace probabilisé E. On dit que

- On dit que X_n converge **en loi** ou (**en distribution**) vers X si pour toutes fonctions f continue bornée sur E on a

$$\mathbb{E}(f(X_n)) \underset{n\to\infty}{\longrightarrow} \mathbb{E}(f(X))$$

  On le note $X_n \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} X$ ou bien $X_n \underset{n\to\infty}{\xrightarrow{d}} X$.

Par magie si les variables aléatoire sont réelles il suffit de montrer que

$$F_n(x) \underset{n\to\infty}{\longrightarrow} F(X) \quad \text{en tout } x \text{ où } F \text{ est continue.}$$

Ici, F_n et F désignes les fonctions de répartition des X_n et de X respectivement.

Loi(s) des grands nombres

Loi forte (admise)

Si $(X_k)_{k\geqslant 1}$ est une suite de *v.a.i.i.d.* telle que $\mathbb{E}[X_1] = m \in \mathbb{R}$, alors

$$\overline{X}_n \xrightarrow[n\to\infty]{p.s.} m.$$

Autrement dit, $\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = m\right) = 1$.

Un estimateur T_n qui converge presque-sûrement vers un paramètre $t \in \mathbb{R}$ est dit **fortement consistant** pour estimer t.

La moyenne empirique $\overline{X}_n$ est fortement consistante pour estimer la moyenne m.

Loi(s) des grands nombres

Loi faible (conséquence de Bienaymé-Tchebychev)

Si $(X_k)_{k\geqslant 1}$ est une suite de *v.a.i.i.d.* telle que $\mathbb{E}[X_1] = m \in \mathbb{R}$, alors

$$\overline{X}_n \xrightarrow[n\to\infty]{\mathbb{P}} m.$$

Autrement dit, pour tout $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\overline{X}_n - m| > \varepsilon\right) = 0$.

Un estimateur T_n qui converge en probabilité vers un paramètre $t \in \mathbb{R}$ est dit **consistant** pour estimer t.

La moyenne empirique $\overline{X}_n$ est consistante pour estimer la moyenne m.

Théorème central limite (admis)

Si $(X_k)_{k\geqslant 1}$ est une suite de *v.a.i.i.d.* telle que $\mathbb{E}[X_1] = m \in \mathbb{R}$ et $\mathrm{Var}(X_1) = \sigma^2 > 0$, alors

$$\sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1).$$

Autrement dit, pour tout $x \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \leqslant x\right) = \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt.$$

Ce théorème fondamental de la théorie des probabilités illustre l'importance de la loi normale. Du point de vue statistique, il permet de manipuler toute moyenne empirique de *v.a.i.i.d.* **correctement normalisée** comme un variable normale dans un cadre asymptotique.

La moyenne empirique $\overline{X}_n$ est dite **asymptotiquement normale**.

Théorème central limite (illustration)

$$Z_n = \sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Question : comment illustrer le résultat d'une convergence en loi ?

En effet, à l'issue d'une simulation ou d'une expérience, nous n'avons à notre disposition qu'**une unique réalisation** de la variable aléatoire qui est l'objet de cette convergence.

Pour l'exemple de la moyenne empirique normalisée Z_n, une fois les réalisations des n *v.a.i.i.d.* $X_1, \ldots, X_n$ générées, nous pouvons calculer la réalisation de Z_n mais pas illustrer sa **loi en tant que variable aléatoire**.

Théorème central limite (illustration)

$$Z_n = \sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Question : comment illustrer le résultat d'une convergence en loi ?

En effet, à l'issue d'une simulation ou d'une expérience, nous n'avons à notre disposition qu'**une unique réalisation** de la variable aléatoire qui est l'objet de cette convergence.

Pour l'exemple de la moyenne empirique normalisée Z_n, une fois les réalisations des n *v.a.i.i.d.* $X_1, \ldots, X_n$ générées, nous pouvons calculer la réalisation de Z_n mais pas illustrer sa **loi en tant que variable aléatoire**.

Nous allons devoir répéter l'expérience m fois pour obtenir des réalisations de variables $Z_n^{(1)}, \ldots, Z_n^{(m)}$ indépendantes et même loi que Z_n.

Théorème central limite (illustration)

$$Z_n = \sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Question : comment illustrer le résultat d'une convergence en loi ?

Une première idée « naïve » consiste à **approcher la densité** de Z_n par un histogramme (ou un estimateur à noyau) tel que nous l'avons présenté dans la partie sur la statistique exploratoire.

Cette méthode est **acceptable d'un point de vue asymptotique** mais présente un inconvénient en pratique : les blocs sont de taille égale et, par construction, ne contiennent **pas le même nombre de données**. De fait, la qualité de l'estimation n'est pas constante dans chaque bloc et la représentation de la densité obtenue peut diverger de celle attendue, en particulier dans les **régions de faible probabilité**.

Théorème central limite (illustration)

$$Z_n = \sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

Question : comment illustrer le résultat d'une convergence en loi ?

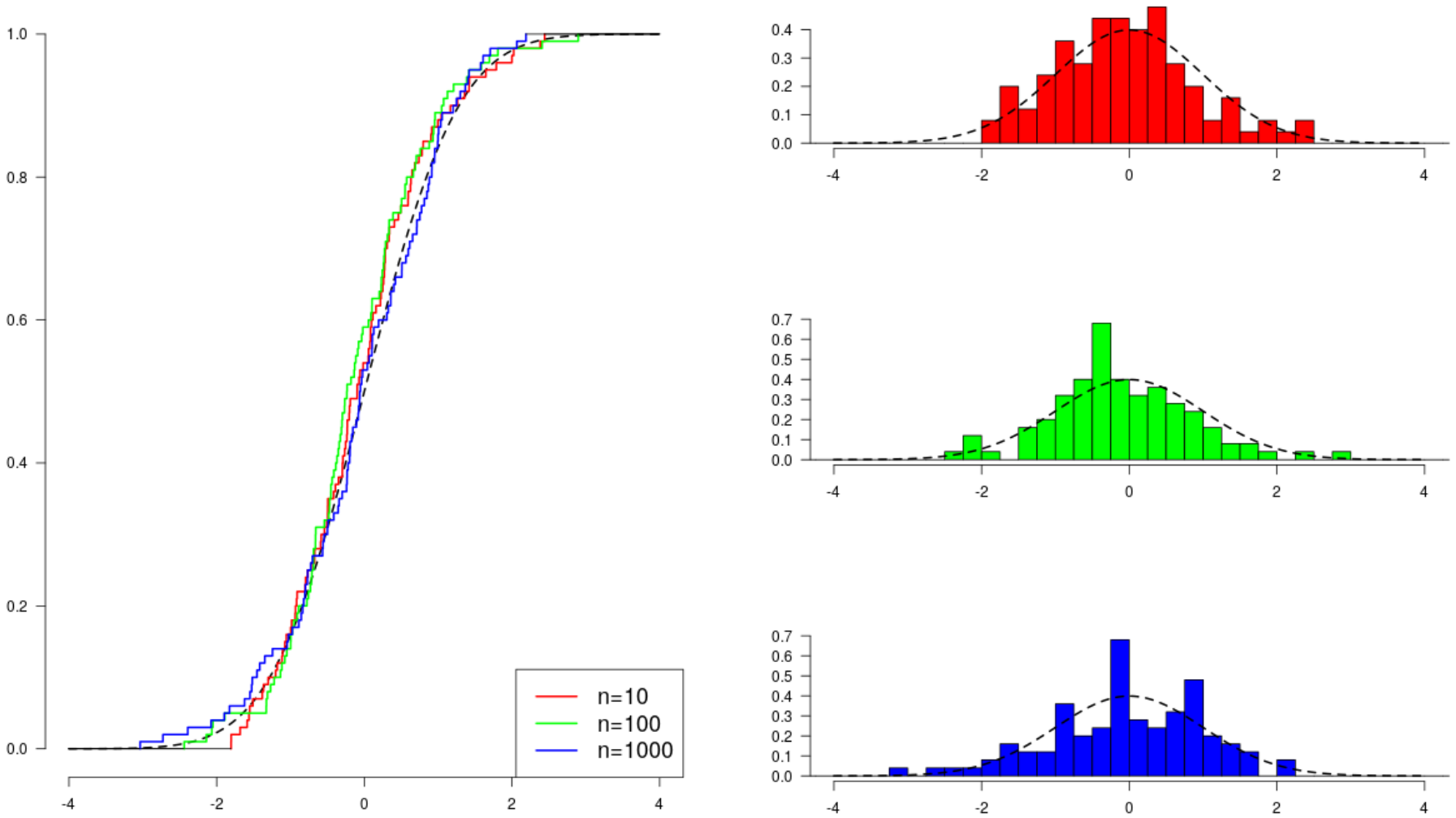
Une méthode alternative est basée sur la **fonction de répartition empirique**,

$$\forall x \in \mathbb{R},\ F_m(x) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}_{Z_n^{(j)} \leqslant x}.$$

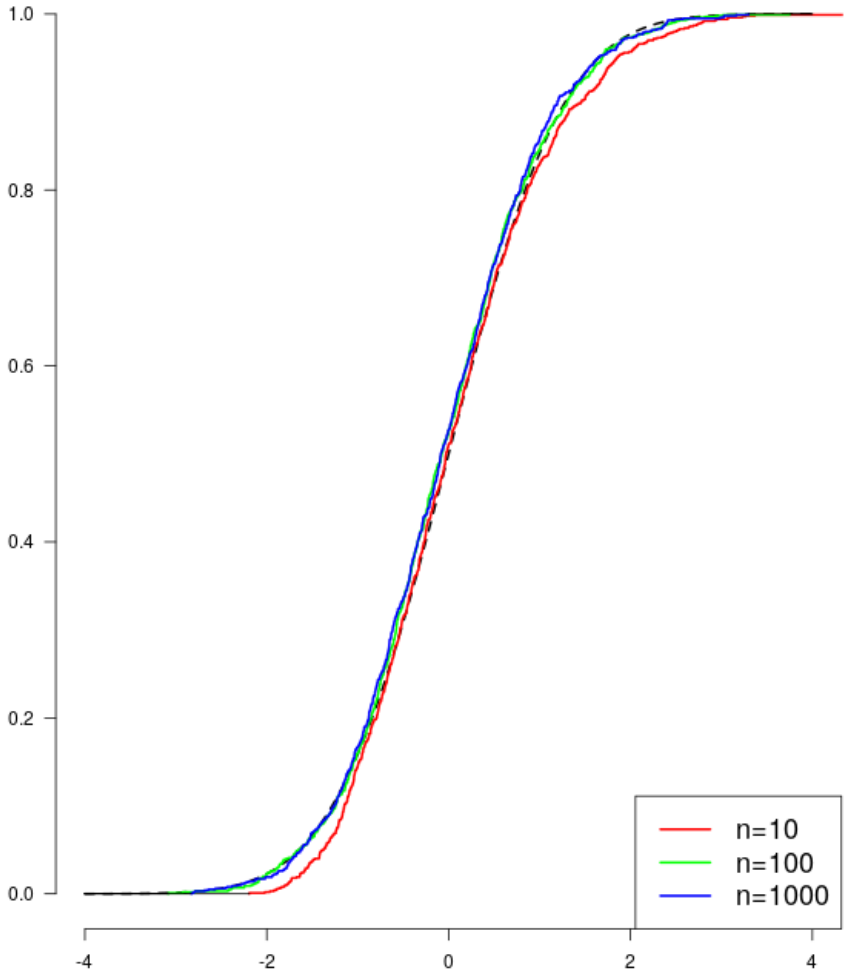
En effet, si F_{Z_n} est la fonction de répartition de la variable aléatoire Z_n, le théorème de Kolmogorov-Smirnov donne la convergence uniforme et presque-sûre de F_m vers F_{Z_n},

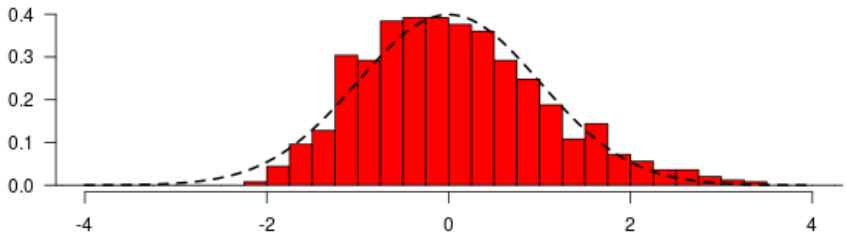
$$\sup_{x\in\mathbb{R}} |F_m(x) - F_{Z_n}(x)| \xrightarrow[m\to\infty]{p.s.} 0.$$

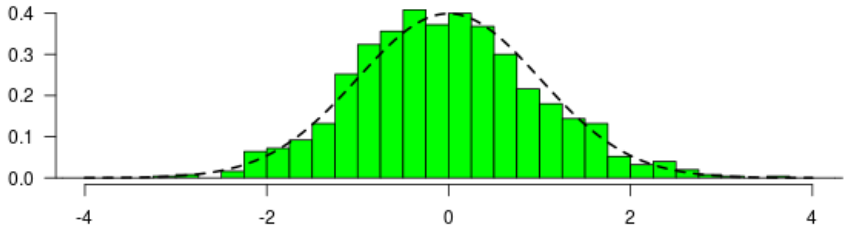
Théorème central limite (illustration, loi $\mathcal{E}(3)$, $m = 100$)

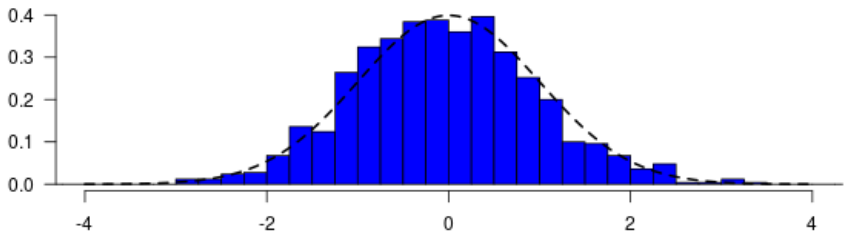


Théorème central limite (illustration, loi $\mathcal{E}(3)$, $m = 1000$)

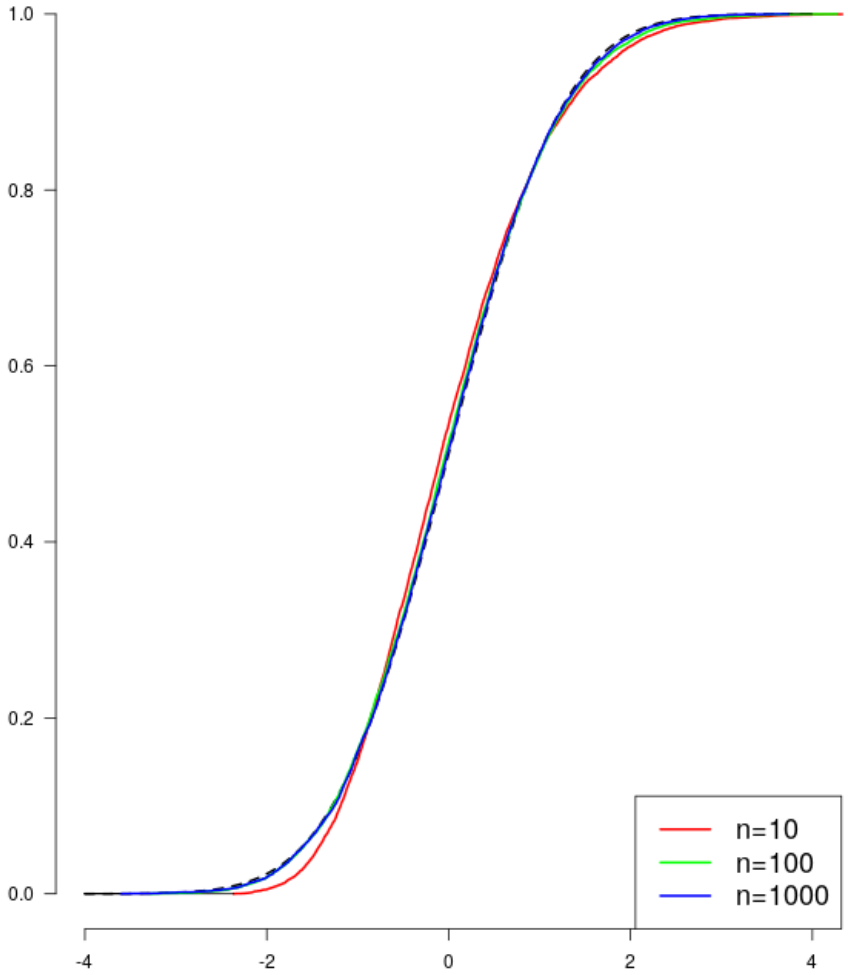


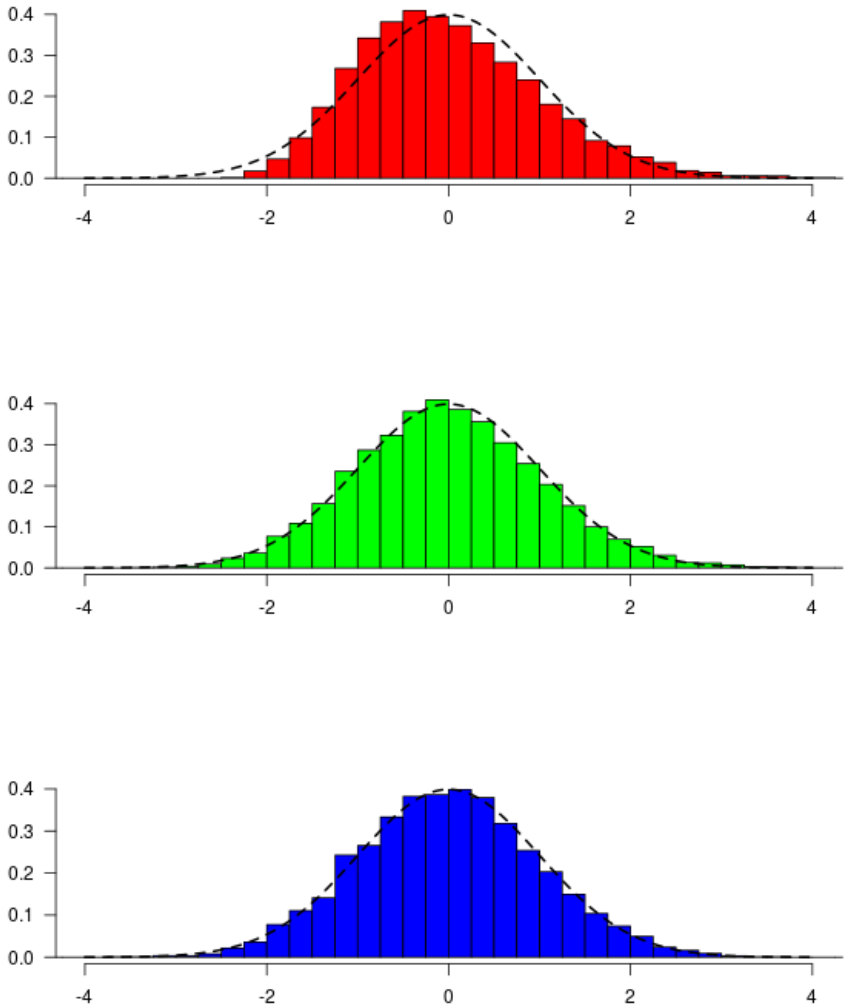






Théorème central limite (illustration, loi $\mathcal{E}(3)$, $m = 10000$)





Intervalle de confiance asymptotique

Dans le cas de *v.a.i.i.d.* $X_1, \dots, X_n$ avec $\mathbb{E}[X_1] = m \in \mathbb{R}$ et $\mathrm{Var}(X_1) = \sigma^2 > 0$, le théorème central limite permet d'écrire que, pour tout $\alpha \in]0,1[$,

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{X}_n - m < \frac{x_{\alpha/2}\sqrt{\sigma^2}}{\sqrt{n}}\right) = \frac{\alpha}{2}$$

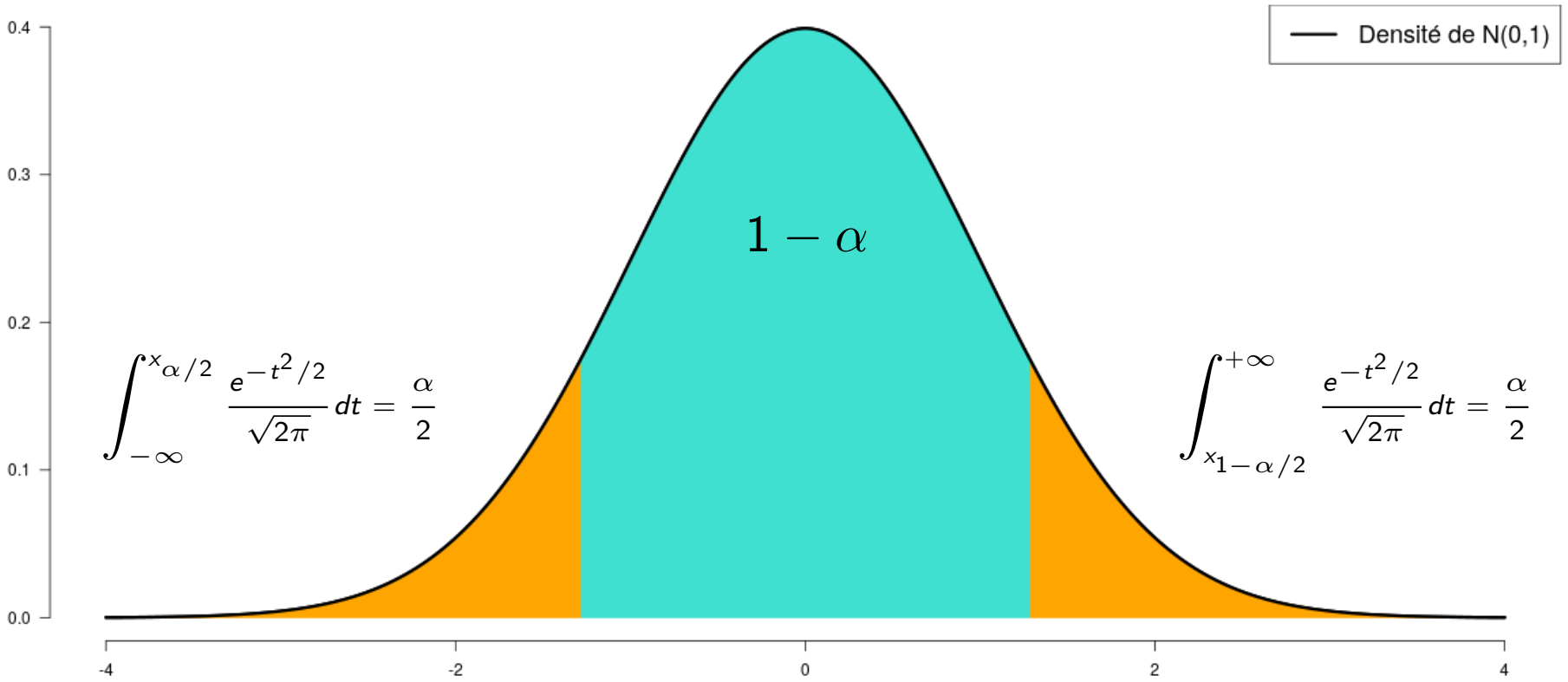
où $x_{\alpha/2} \in \mathbb{R}$ est le **quantile** d'ordre $\alpha/2$ de la loi normale centrée réduite,

$$\int_{-\infty}^{x_{\alpha/2}} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt = \frac{\alpha}{2}.$$

De même,

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{X}_n - m > \frac{x_{1-\alpha/2}\sqrt{\sigma^2}}{\sqrt{n}}\right) = \frac{\alpha}{2}.$$

Intervalle de confiance asymptotique



$$x_{1-\alpha/2} = -x_{\alpha/2}$$

Intervalle de confiance asymptotique

Nous obtenons finalement que

$$\lim_{n\to\infty} \mathbb{P}\left(m \in \left]\overline{X}_n - \frac{x_{1-\alpha/2}\sqrt{\sigma^2}}{\sqrt{n}}; \overline{X}_n + \frac{x_{1-\alpha/2}\sqrt{\sigma^2}}{\sqrt{n}}\right[\right) = 1-\alpha.$$

Si la variance σ^2 est **connue**, il s'agit d'un **intervalle de confiance asymptotique** de niveau $1-\alpha \in]0,1[$.
Si la variance σ^2 est **inconnue** mais peut être estimée de façon **consistante** par un estimateur $\hat{\sigma}_n^2$,

$$\hat{\sigma}_n^2 \xrightarrow[n\to\infty]{\mathbb{P}} \sigma^2,$$

alors il est possible de montrer (*cf.* Lemme de Slutsky) que

$$\left]\overline{X}_n - \frac{x_{1-\alpha/2}\sqrt{\hat{\sigma}_n^2}}{\sqrt{n}}; \overline{X}_n + \frac{x_{1-\alpha/2}\sqrt{\hat{\sigma}_n^2}}{\sqrt{n}}\right[$$

est encore un intervalle de confiance asymptotique de niveau $1-\alpha \in]0,1[$.

TCL versus Bienaymé-Tchebychev

Avec les mêmes arguments, si l'estimateur $\hat{\sigma}_n^2$ de la variance est consistant, l'intervalle IC_3 obtenu précédemment avec Bienaymé-Tchebychev est également asymptotique de niveau $1 - \alpha \in]0, 1[$.

Comment choisir entre IC_3 et l'intervalle de confiance obtenu avec le théorème central limite ?

TCL versus Bienaymé-Tchebychev

Avec les mêmes arguments, si l'estimateur $\hat{\sigma}_n^2$ de la variance est consistant, l'intervalle IC_3 obtenu précédemment avec Bienaymé-Tchebychev est également asymptotique de niveau $1-\alpha \in]0,1[$.

Comment choisir entre IC_3 et l'intervalle de confiance obtenu avec le théorème central limite ?

Nous pouvons comparer les longueurs de ces intervalles pour choisir le plus « précis » :

- Bienaymé-Tchebychev : $2\sqrt{\frac{\hat{\sigma}_n^2}{n}} \times \frac{1}{\sqrt{\alpha}}$
- TCL : $2\sqrt{\frac{\hat{\sigma}_n^2}{n}} \times x_{1-\alpha/2} \leqslant 2\sqrt{\frac{\hat{\sigma}_n^2}{n}} \times \sqrt{2\ln(2/\alpha)}$ (*cf.* Borne de Chernoff)

Pour α « proche » de 0, l'intervalle de confiance déduit du théorème central limite est donc bien plus court que celui issu de Bienaymé-Tchebychev.

Propriétés d'estimateurs

Lorsqu'on veut estimer un paramètre θ sur un échantillon $X_1, \ldots, X_n$ avec un estimateur T_n, idéalement on aimerai que l'estimateur ai les propriétés suivantes :

- $T_n \underset{n\to\infty}{\longrightarrow} \theta$-p.s. On dit alors qu'il est **asymptotiquement sans biais** ou **fortement consistent**.
- $\sqrt{n}(T_n - \theta) \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, \sigma^2)$, on dit alors qu'il est **asymptotiquement normal**, avec σ connue ou que l'on sais estimer.

On connait déjà un estimateur qui à ces deux propriétés ! ! La moyenne empirique grace à la **loi des grands nombres** et au **TCL**.

Théorèmes importants : Continuous mapping

Continuous Mapping Theorem

Soit $(X_n)_{n\in\mathbb{N}} \subset E$ une suite de variable aléatoires et $X \in E$ une variable aléatoire telle que

$$X_n \underset{n\to\infty}{\longrightarrow} X \quad \text{p.s, en proba, en loi.}$$

Si g est une fonction continue sur E alors

$$g(X_n) \underset{n\to\infty}{\longrightarrow} g(X) \quad \text{p.s, en proba, en loi.}$$

Théorèmes importants : delta-method

Delta-Method

Soit $(X_n)_{n\in\mathbb{N}} \subset \mathbb{R}$ une suite de variable aléatoires et $\theta \in \mathbb{R}$ telle que

$$\sqrt{n}(X_n - \theta) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, \sigma^2),$$

avec $var(X_n)) = \sigma^2$. Si g est une fonction C^1 sur E alors

$$\sqrt{n}(g(X_n) - (\theta)) \underset{n\to\infty}{\longrightarrow} \mathcal{N}(0, (g'(\theta)\sigma)^2).$$

Il existe des versions (y compris plus faible) de ce théorème sur $\mathbb{R}^d$ ou sur des espaces plus compliqués. Si vous êtes intéressé allez voir "Asymptotic Statistics" de P. Billingsley.

Méthode des moments

Soit $X_1, \ldots, X_n$ un échantillon i.i.d sur lequel on veut estimer un paramètre θ. Supposons qu'il existe une fonction f, inversible, telle que $\mathbb{E}(X_1) = f(\theta)$.

Idée de la méthode des moments

- On sait que $\bar{X}_n$ est un bon estimateur de $\mathbb{E}(X_1)$.
- Comme f est inversible, on a $f^{-1}(\mathbb{E}(X_1)) = \theta$.
- Naturellement on veut estimer θ par

$$\hat{\theta}_n^{MM} = f^{-1}(\bar{X}_n).$$

- Si f^{-1} est continue alors $\hat{\theta}_n^{MM}$ est fortement consitent. (Loi des grands nombres et continuous mapping theorem).
- Si f^{-1} est C^1 alors $\hat{\theta}_n^{MM}$ est asymptotiquement normal. (TCL et Delta-Method).

Maximum de Vraisemblance

Soit $X_1, \ldots, X_n$ un échantillon i.i.d sur lequel on veut estimer un paramètre θ. On défnit

$$p(x;\theta) = \begin{cases} \mathbb{P}_\theta(X_1 = x) & \text{dans le cas discret,} \\ f_\theta(x) & \text{dans le cas continu,} \end{cases}$$

où f_θ désigne la densité de la loi $\mathbb{P}_\theta$ si elle est absolument continue.

La vraisemblance

$$\mathcal{L}(\theta, X_1, \ldots, X_n) = \prod_{i=1}^{n} p(X_i, \theta).$$

L'estimateur du maximum de vraisemblance est alors

$$\hat{\theta}_n^{MV} = \arg\max_\theta \mathcal{L}(\theta; \mathbf{X}_n).$$

Dans la pratique on s'intéressera plus souvent à $log(\mathcal{L}(\theta, X_1, \ldots, X_n))$, ce qui facilitera les calculs.