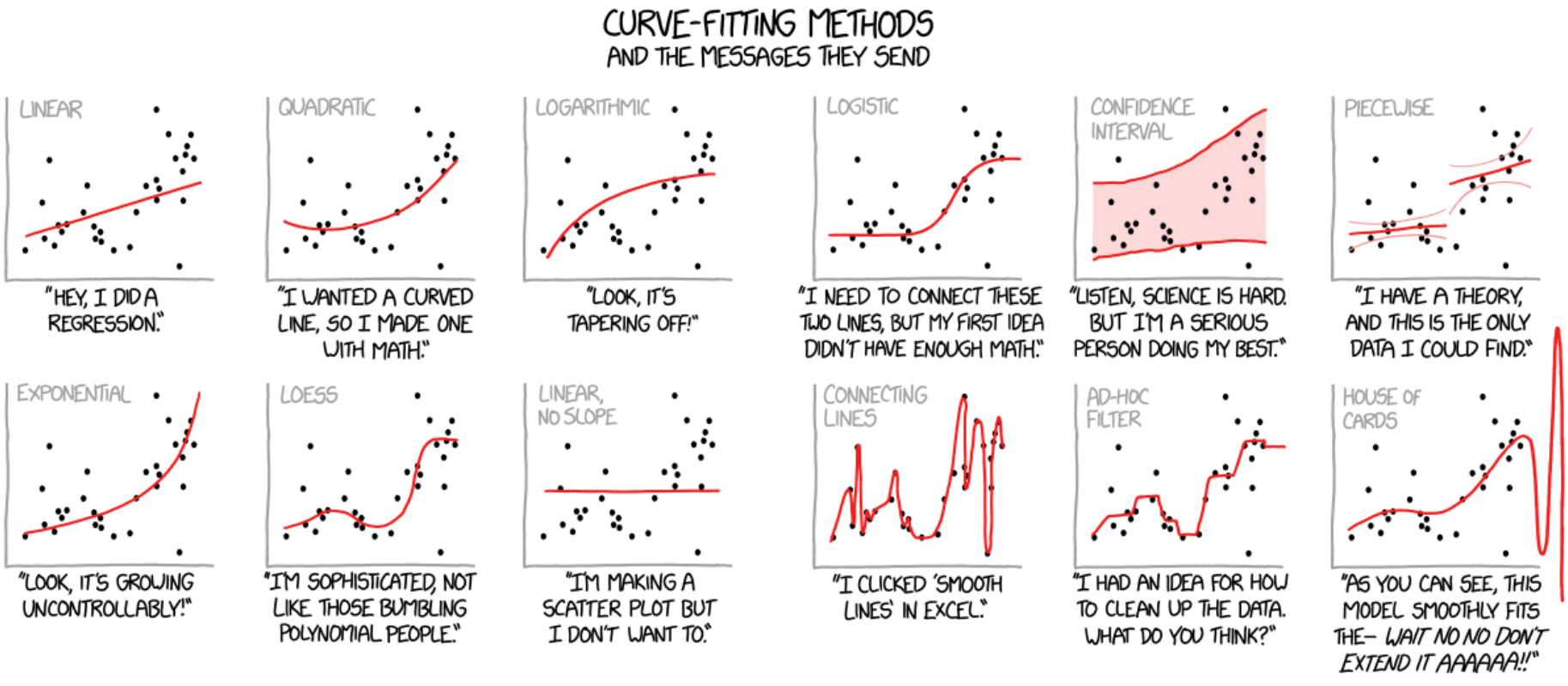
Statistique

Benjamin Bobbia

ISAE

Statistique inférentielle



Curve-Fitting, XKCD, xkcd.com/2048

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 $X_1, \dots, X_n$, l'échantillon est constitué de valeurs observées $x_1, \dots, x_n \in \mathbb{R}$.

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 $X_1, \dots, X_n$ 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 $X_1, \dots, X_n$ 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,

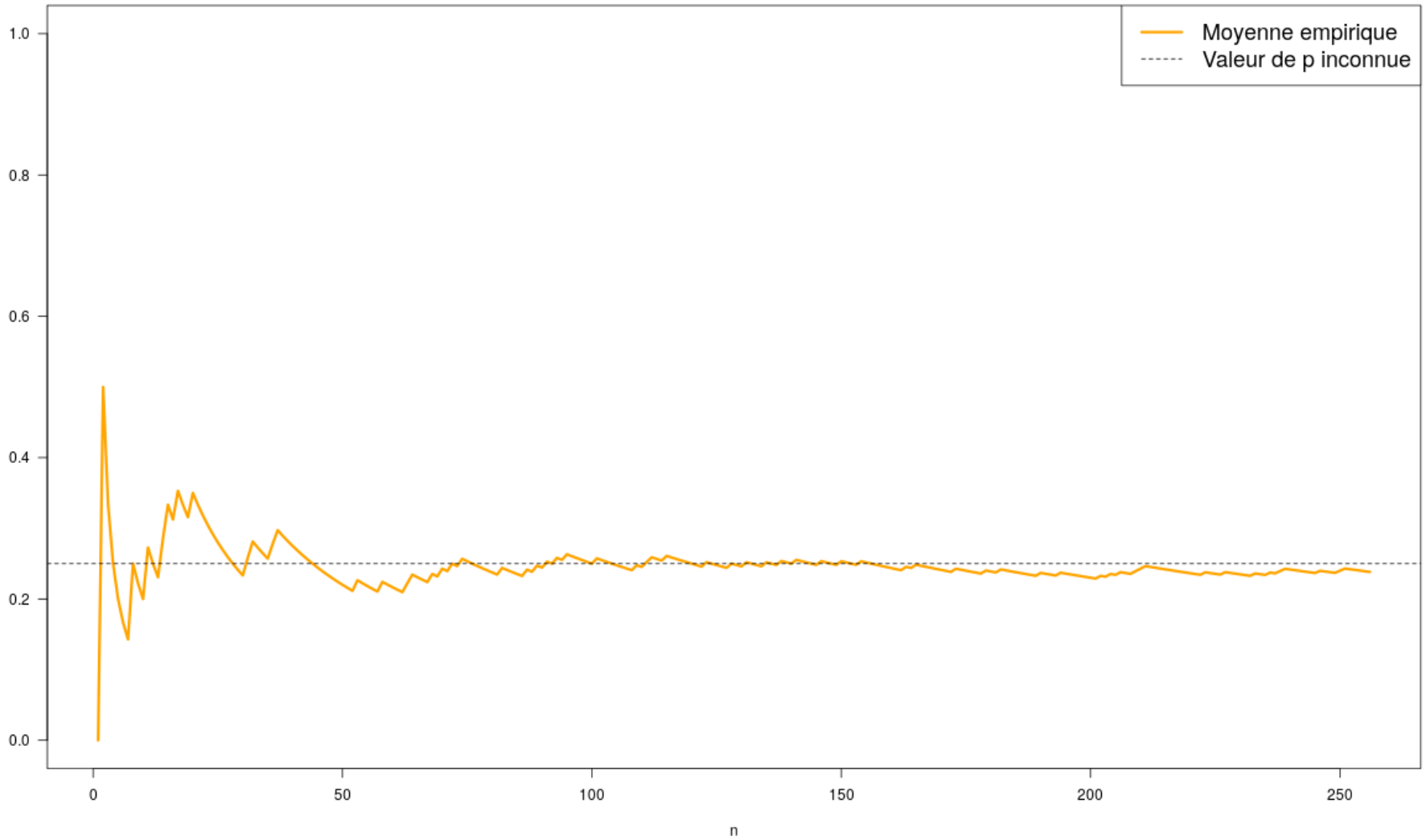
$$\forall k \in \{1,\dots,n\},\ X_k = \begin{cases} 1 & \text{si le } k\text{ème individu présente la mutation,} \\ 0 & \text{sinon.} \end{cases}$$

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

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

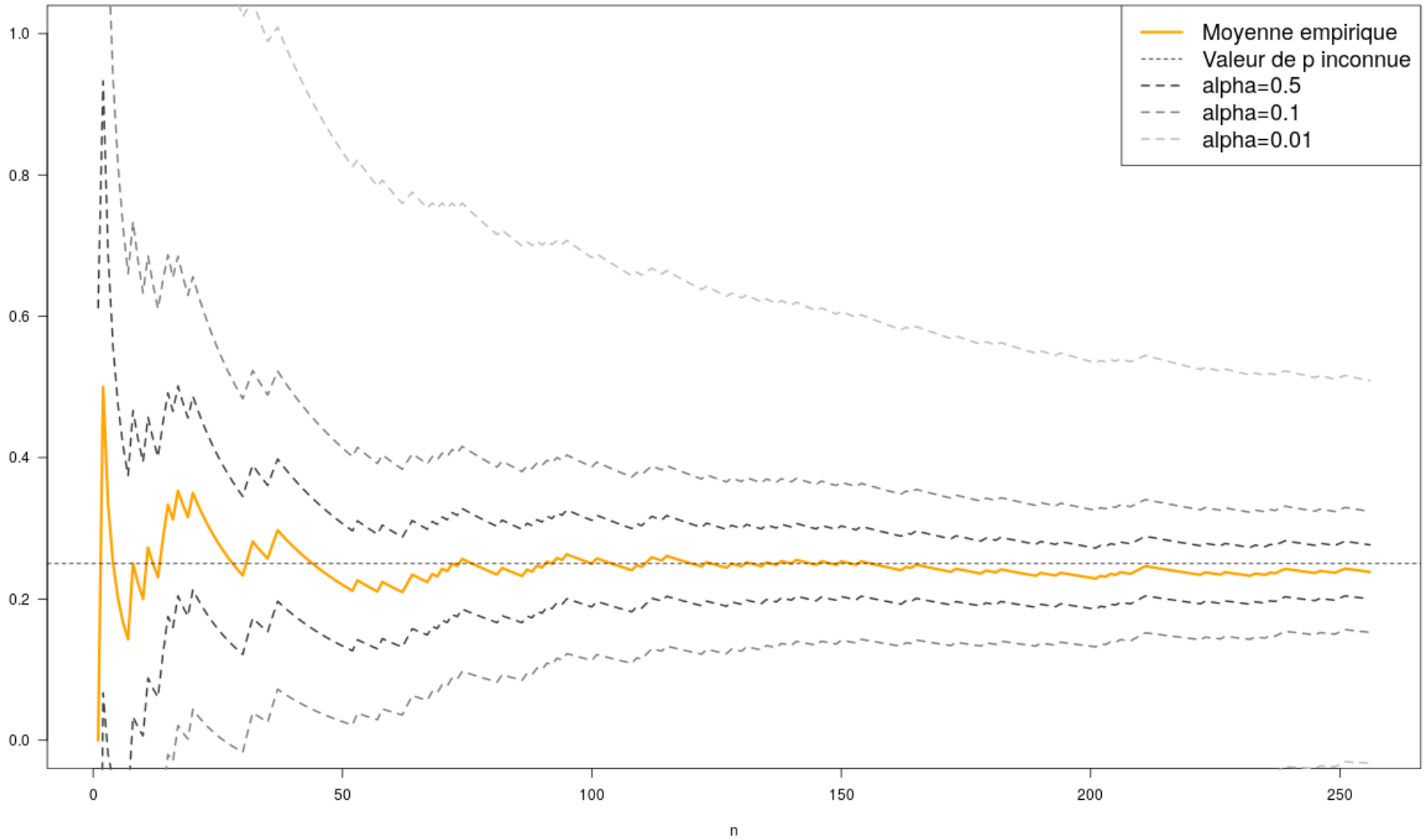
car $\mathrm{Var}(X_1) = p(1-p)$.

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.

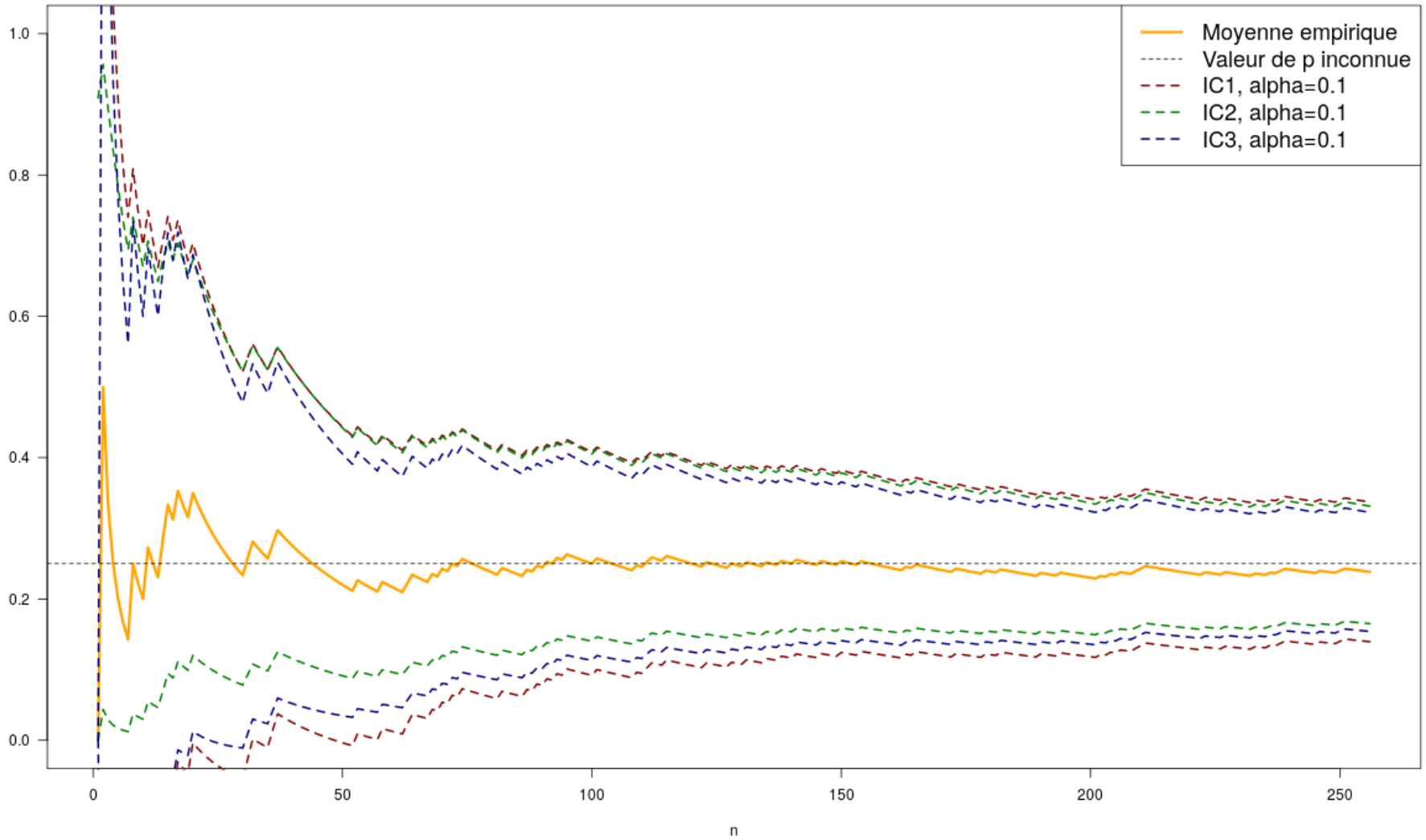
Si nous disposons d'un **estimateur de la variance** $\hat{\sigma}_n^2$, il est possible de l'utiliser pour définir l'intervalle

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

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 ?

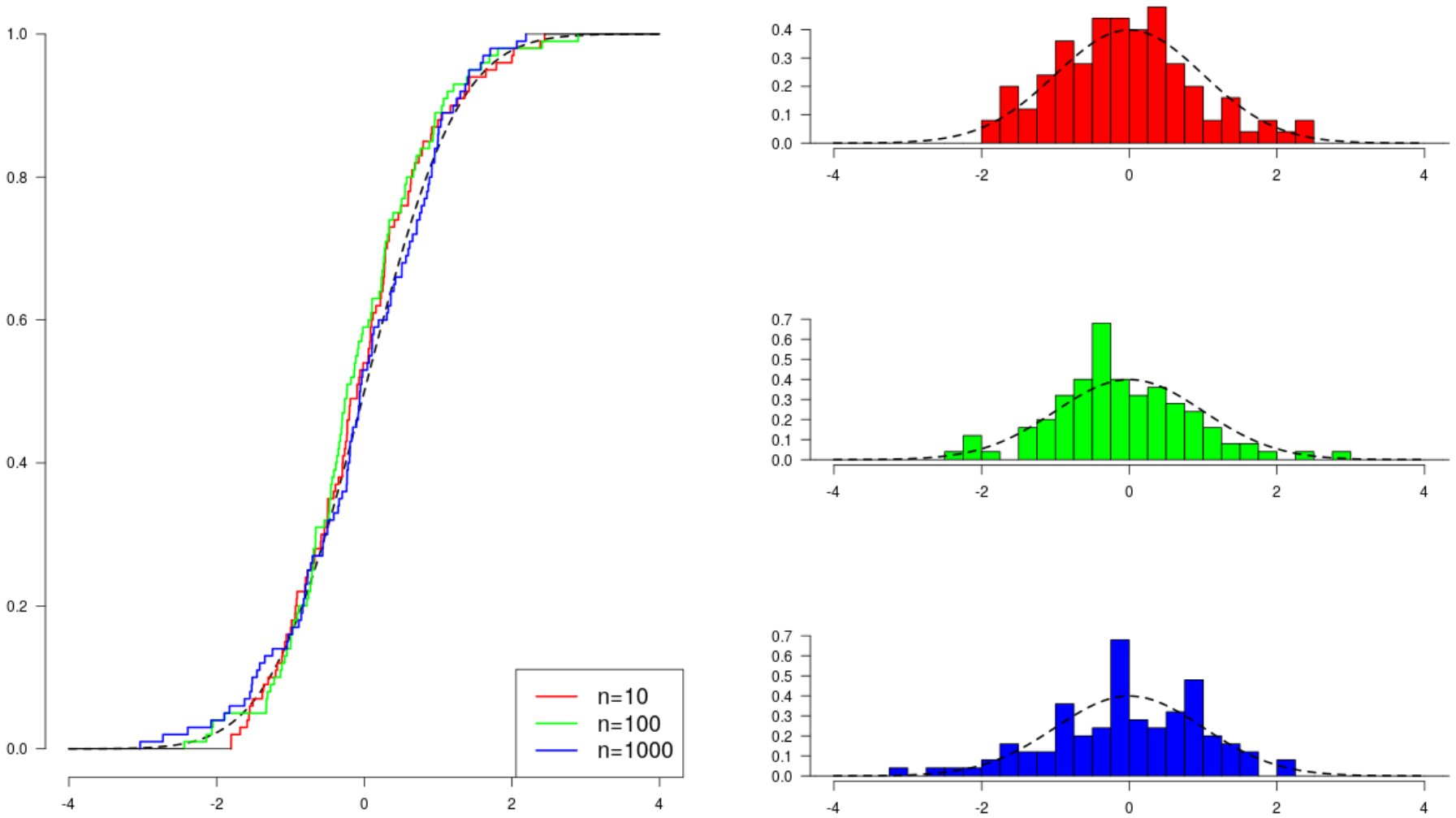
Une méthode alternative est basée sur la **fonction de répartition empirique**,

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

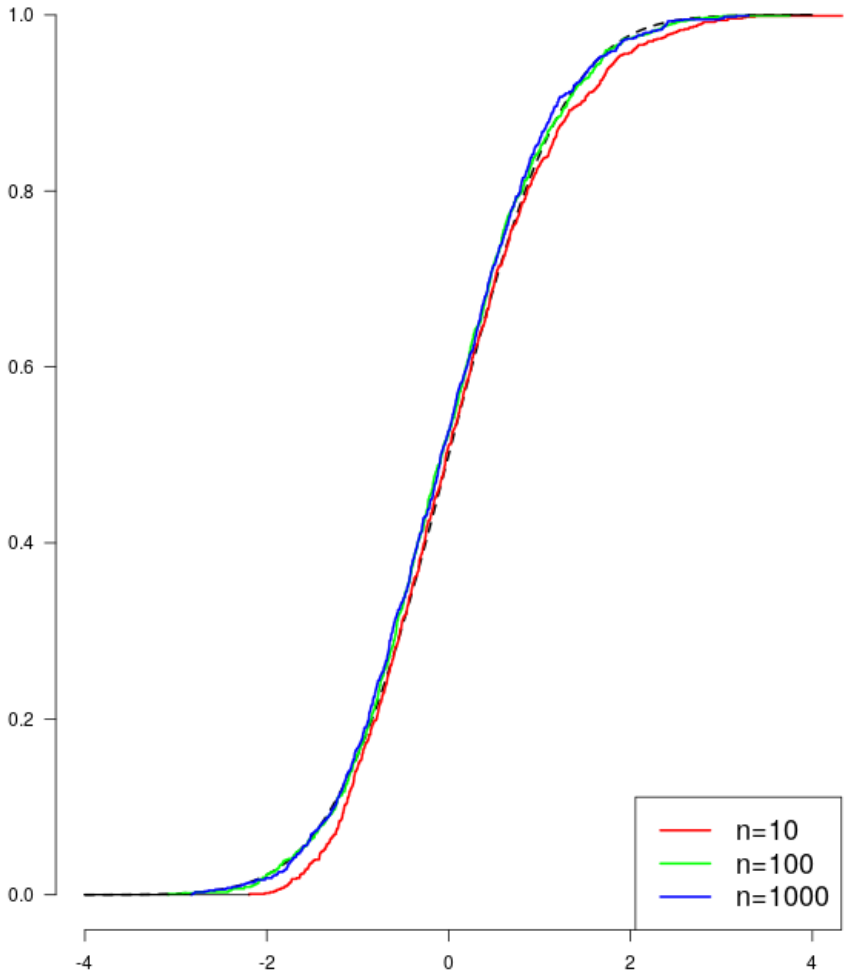
En effet, si F_{Z_n} est la fonction de répartition de la variable aléatoire Z_n, le théorème de Kolmogorov-Smirnov donne la convergence uniforme et presque-sûre de F_m vers F_{Z_n},

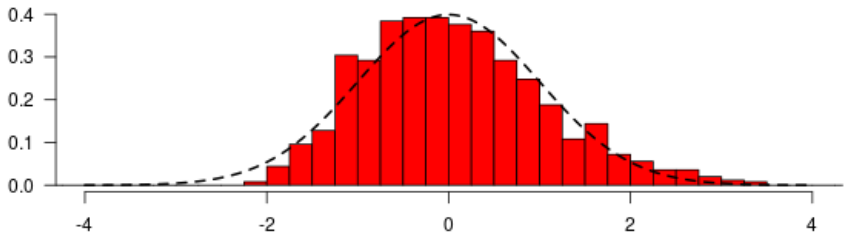
$$\sup_{x\in\mathbb{R}} |F_m(x) - F_{Z_n}(x)| \xrightarrow[m\to\infty]{p.s.} 0.$$

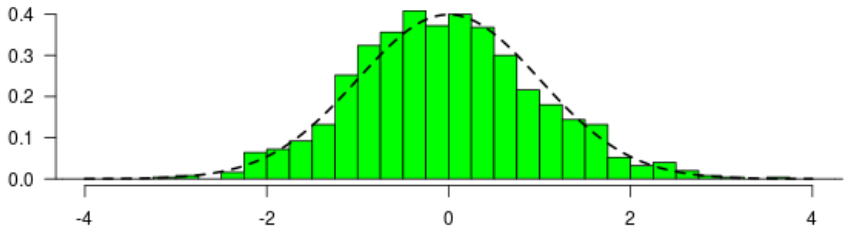
Théorème central limite (illustration, loi $\mathcal{E}(3)$, $m = 100$)

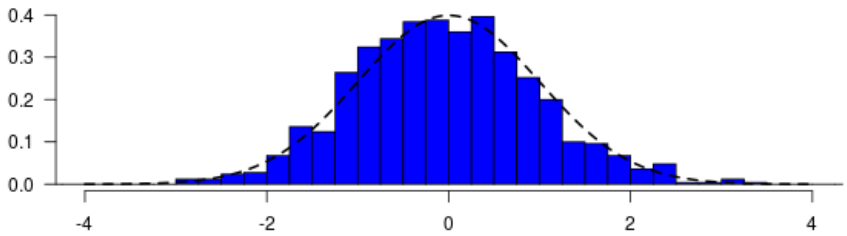


Théorème central limite (illustration, loi $\mathcal{E}(3)$, $m = 1000$)

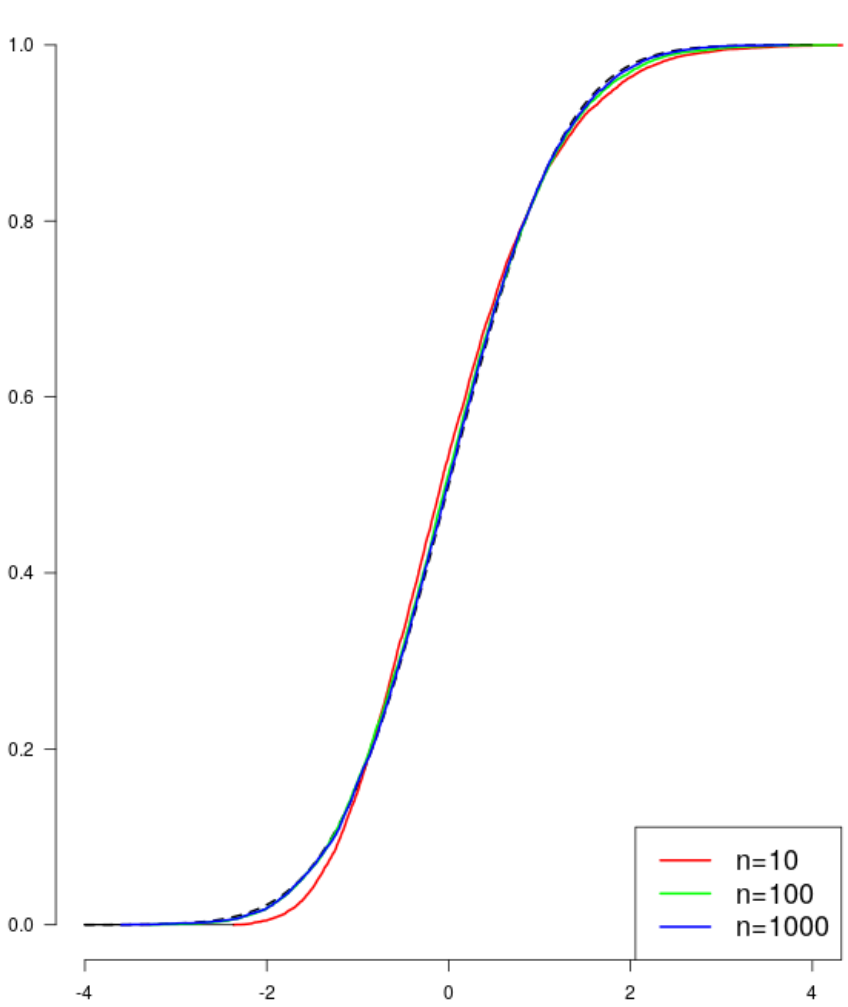


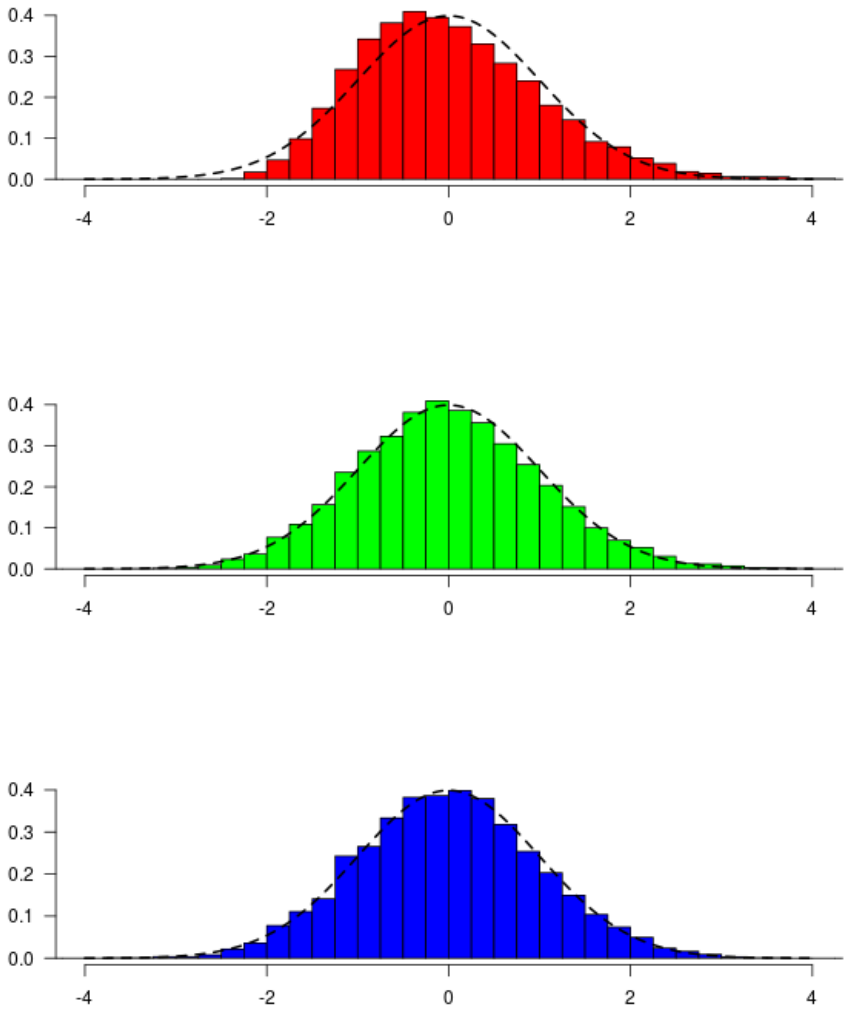






Théorème central limite (illustration, loi $\mathcal{E}(3)$, $m = 10000$)





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 $\alpha \in]0,1[$,

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

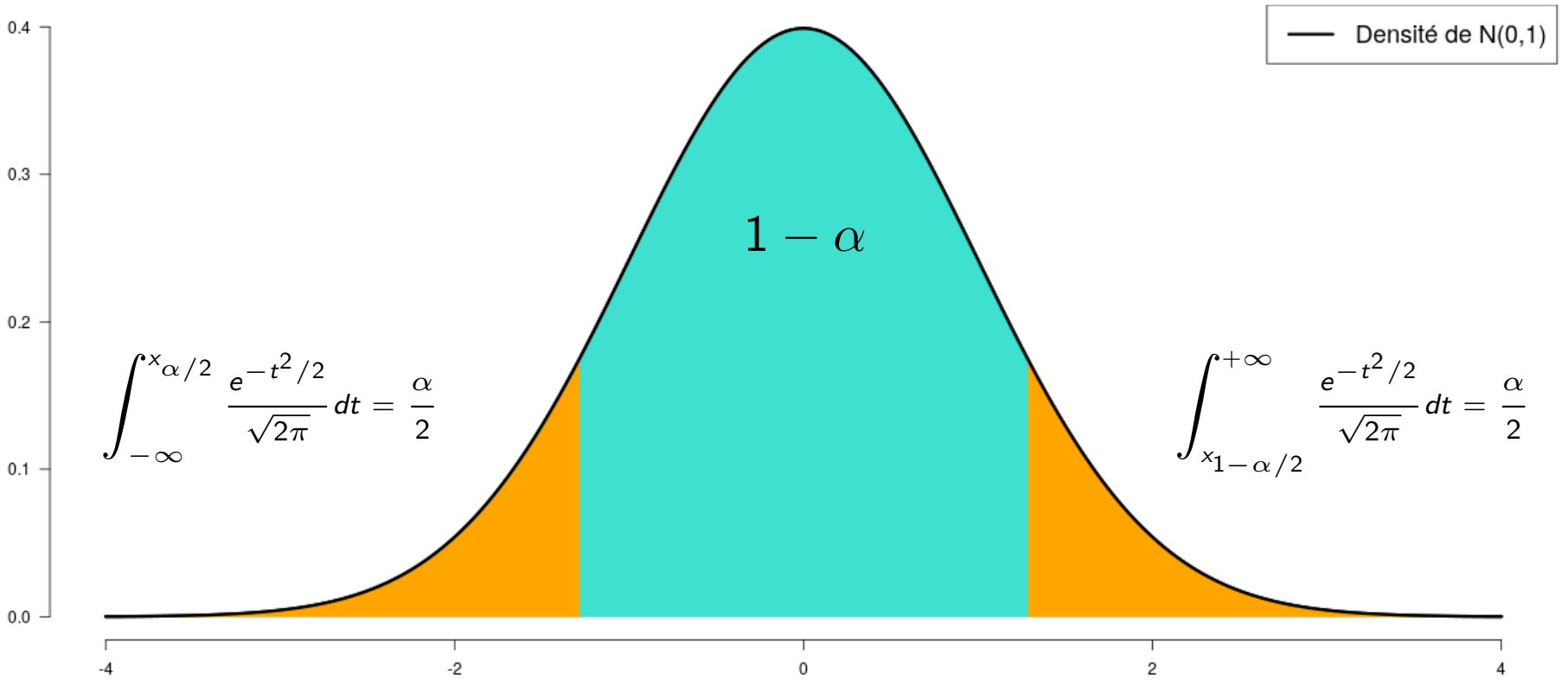
où $x_{\alpha/2} \in \mathbb{R}$ est le **quantile** d'ordre $\alpha/2$ de la loi normale centrée réduite,

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

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 $X_1, \dots, X_n$ 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.

3.0 Tests statistiques

Motivation à partir d'un intervalle de confiance

Reprenons l'étude de la proportion $p \in]0,1[$ d'une population qui présente une mutation génétique particulière. La généticienne a de bonnes raisons de penser que la moitié de la population a cette mutation dans son génome. Elle souhaite donc **valider** cette hypothèse « $p = 1/2$ ». Si elle est amenée à **rejeter** son hypothèse de travail, elle en conclura que la mutation est sur-représentée ($p > 1/2$) ou sous-représentée ($p < 1/2$), *i.e.* « $p \neq 1/2$ ».

En termes statistiques, nous disons qu'elle veut **tester** l'**hypothèse nulle**

$$H_0 \ : \ p = \frac{1}{2}$$

contre l'**hypothèse alternative**

$$H_1 \ : \ p \neq \frac{1}{2}.$$

Motivation à partir d'un intervalle de confiance

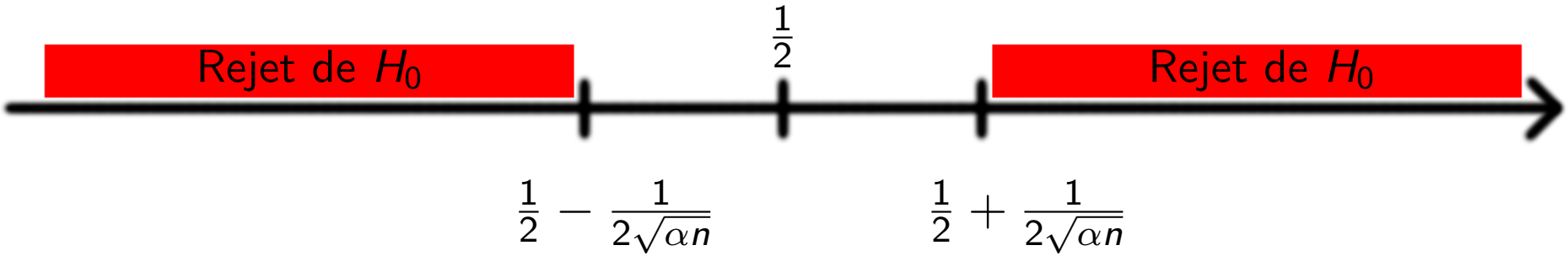
Grâce à son étude statistique précédente effectuée sur n individus tirés uniformément avec remise, elle dispose de l'estimateur $\overline{X}_n$ de p et de l'intervalle de confiance de niveau $1-\alpha \in]0,1[$ donné par Bienaymé-Tchebychev,

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

Elle décide donc d'**accepter l'hypothèse nulle** si

$$\left|\overline{X}_n - \frac{1}{2}\right| < \frac{1}{2\sqrt{\alpha n}}$$

et de **rejeter l'hypothèse nulle** si ce n'est pas le cas.



Motivation à partir d'un intervalle de confiance

Si l'hypothèse nulle H_0 est vraie, alors $p = 1/2$ et la probabilité de prendre la bonne décision est donnée par

$$\mathbb{P}_{H_0}\left(\frac{1}{2} \in \left]\overline{X}_n - \frac{1}{2\sqrt{\alpha n}}; \overline{X}_n + \frac{1}{2\sqrt{\alpha n}}\right[\right) \geqslant 1 - \alpha.$$

Autrement dit, la généticienne fera une **erreur** en rejetant l'hypothèse H_0 si elle est vraie avec une probabilité inférieure à α.

Si l'hypothèse alternative H_1 est vraie, alors $p \neq 1/2$ et il faudrait idéalement rejeter l'hypothèse nulle. Cela a lieu avec une probabilité donnée par

$$\begin{aligned}\mathbb{P}_{H_1}&\left(\frac{1}{2} \notin \left]\overline{X}_n - \frac{1}{2\sqrt{\alpha n}}; \overline{X}_n + \frac{1}{2\sqrt{\alpha n}}\right[\right)\\ &= 1 - \mathbb{P}_{H_1}\left(\frac{1}{2} - \frac{1}{2\sqrt{\alpha n}} \leqslant \overline{X}_n \leqslant \frac{1}{2} + \frac{1}{2\sqrt{\alpha n}}\right).\end{aligned}$$

Motivation à partir d'un intervalle de confiance

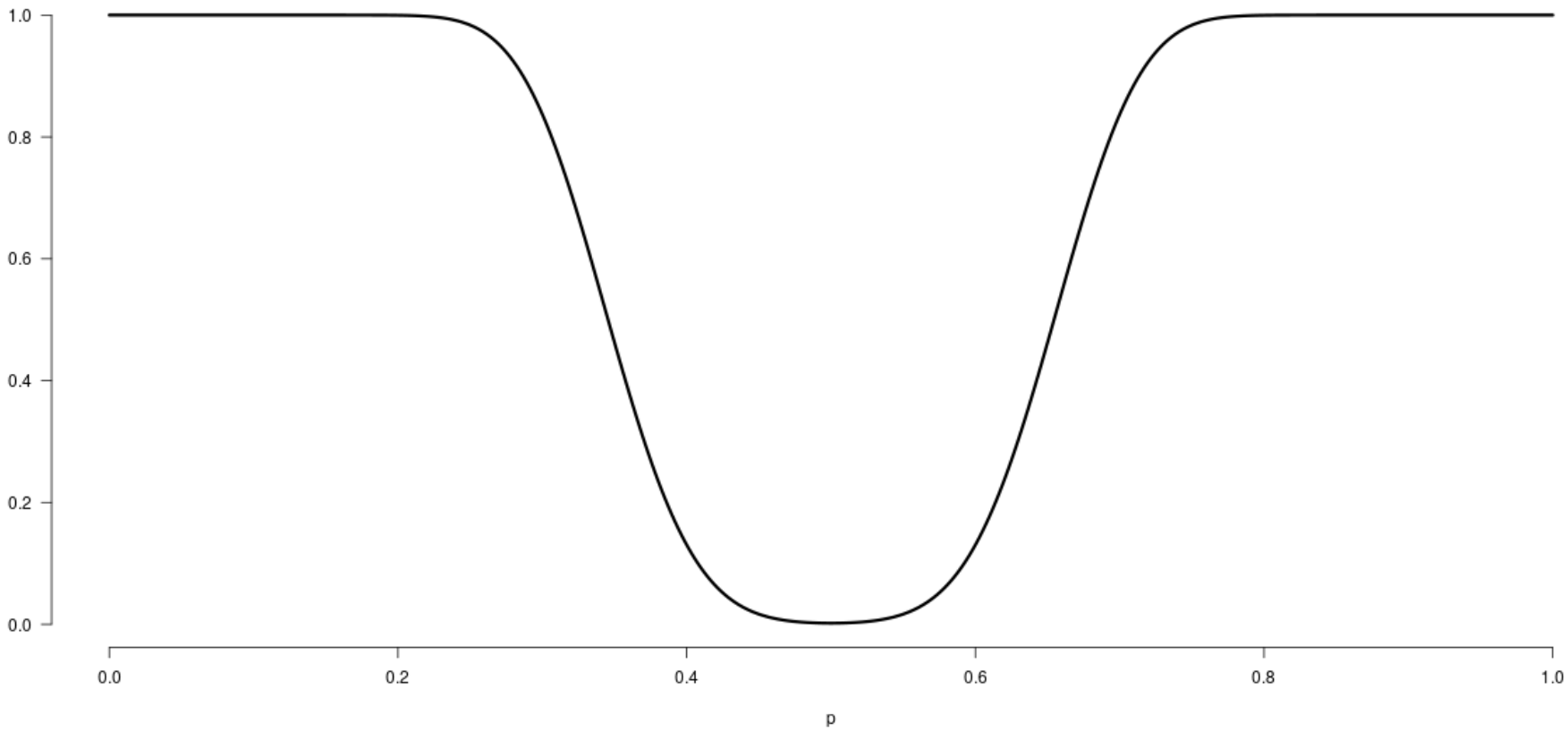
Par définition, $n\overline{X}_n$ est le nombre d'individus présentant la mutation génétique parmi les n individus tirés uniformément avec remise. La loi de cette variable aléatoire à valeurs dans $\{0, \dots, n\}$ est une loi binomiale $\mathcal{B}(n, p)$.

En notant $F_{n,p}$ la fonction de répartition de la loi $\mathcal{B}(n, p)$, nous obtenons que la probabilité de **rejeter l'hypothèse nulle H_0 si l'hypothèse alternative H_1 est vraie** vaut

$$\begin{aligned}\mathbb{P}_{H_1}\left(\frac{1}{2} \notin \left]\overline{X}_n - \frac{1}{2\sqrt{\alpha n}}; \overline{X}_n + \frac{1}{2\sqrt{\alpha n}}\right[\right) \\ = 1 - F_{n,p}\left(\frac{n}{2} + \frac{\sqrt{n}}{2\sqrt{\alpha}}\right) + F_{n,p}\left(\frac{n}{2} - \frac{\sqrt{n}}{2\sqrt{\alpha}}\right).\end{aligned}$$

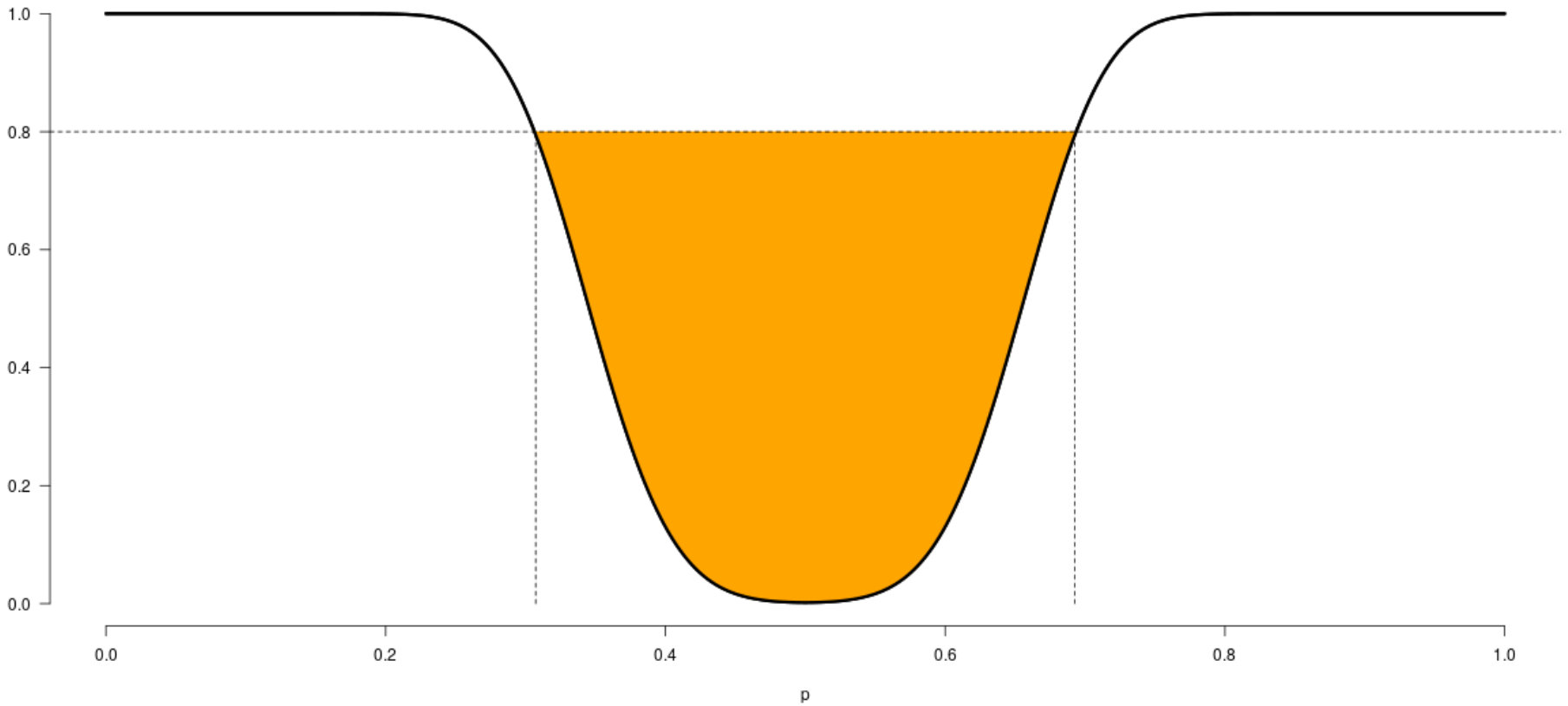
Il s'agit d'une fonction de p.

Motivation à partir d'un intervalle de confiance



$$p \in]0,1[\longmapsto 1 - F_{100,p}\left(\frac{100}{2} + \frac{\sqrt{100}}{2\sqrt{0.1}}\right) + F_{100,p}\left(\frac{100}{2} - \frac{\sqrt{100}}{2\sqrt{0.1}}\right)$$

Motivation à partir d'un intervalle de confiance



La probabilité de rejeter l'hypothèse nulle H_0 lorsque l'hypothèse alternative H_1 est vraie dépend de p et est d'autant plus grande que p est « loin » de $1/2$. Cette fonction est appelée la **puissance** du test.

Principe d'un test statistique

1. Définir une **hypothèse nulle** H_0 et une **hypothèse alternative** H_1.
2. Choisir un **niveau de confiance** $1 - \alpha \in]0, 1[$.
3. Proposer une **règle de décision** telle que

$$\mathbb{P}_{H_0}\left(\text{« Rejeter } H_0 \text{ »}\right) \leqslant \alpha.$$

4. Appliquer la règle de décision avec les **valeurs observées** dans l'échantillon.
5. Conclure si nous acceptons ou rejetons l'hypothèse H_0.

Remarque : pour deux tests de H_0 contre H_1 même niveau de confiance, il est préférable de choisir celui qui a tendance à avoir la plus grande fonction de puissance. Ce n'est pas un ordre total ...

Différentes erreurs

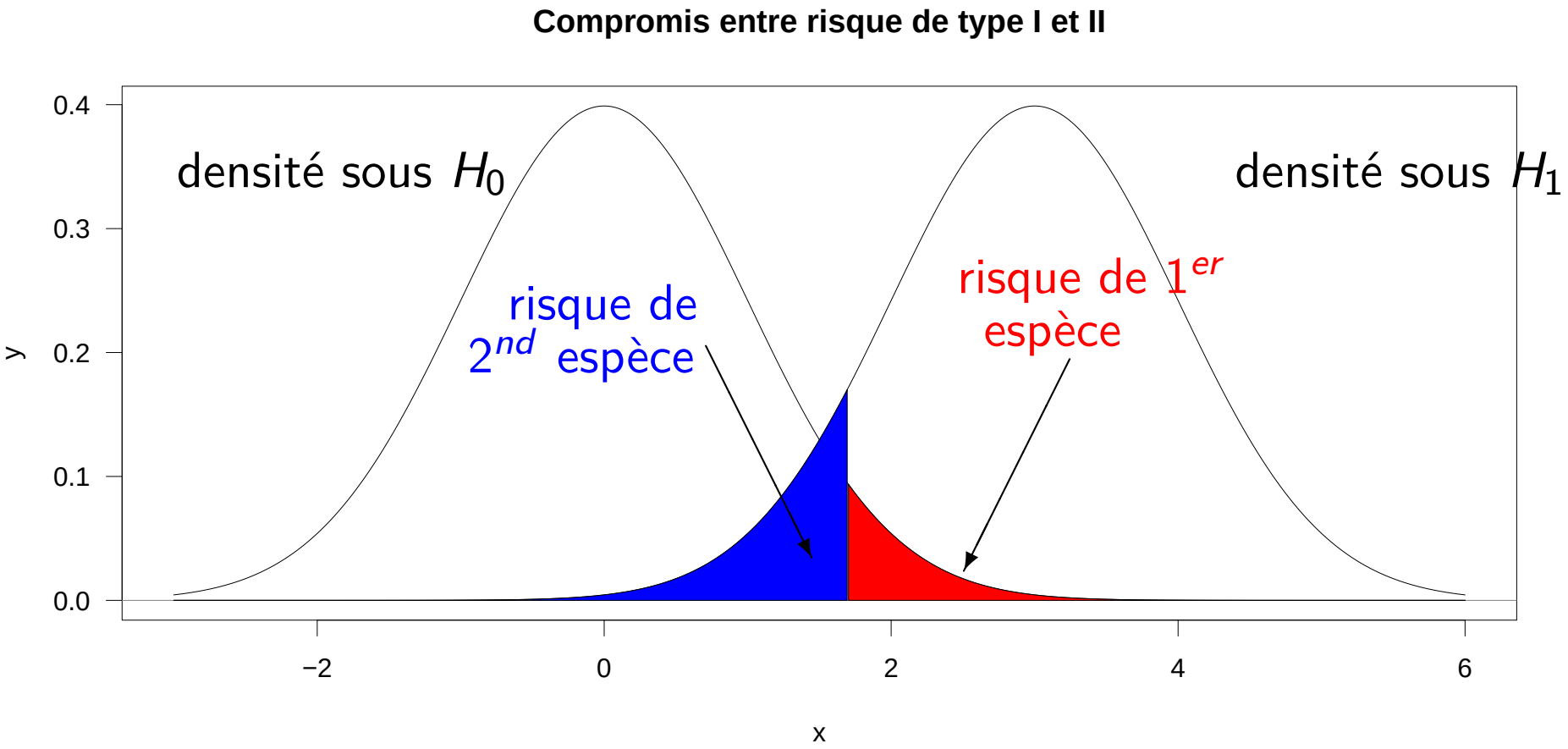
	Accepter H_0	Rejeter H_0
H_0 est vraie	Bonne décision	**Erreur de 1ère espèce**
H_1 est vraie	**Erreur de 2ème espèce**	Bonne décision

Exemple (extrême) : une étude médicale doit être menée pour valider la non dangerosité d'un nouveau médicament proposé par un laboratoire médical. L'utilisation d'un test de l'hypothèse nulle H_0 : « Le médicament est mortel » contre l'hypothèse alternative H_1 : « Le médicament n'est pas mortel » peut conduire aux deux erreurs suivantes :

- **1ère espèce :** médicament déclaré sain alors qu'il est mortel.
  $\Rightarrow$ C'est grave, des gens vont mourir !
- **2ème espèce :** médicament déclaré mortel alors qu'il est sans danger.
  $\Rightarrow$ C'est dommage pour le laboratoire mais personne ne va mourir.

Différentes erreurs, un compromis (encore...)

Il est **impossible** d'avoir à la fois un risque de 1ère espère **et** de 2ème espèce faible.



Exemple du pain d'épice

Une usine agroalimentaire produit du pain d'épice en tranches. Un des critères de qualité est que l'angle de rupture d'une tranche doit être supérieur à 42°. De nombreux facteurs (humidité ambiante, dosage des ingrédients, ...) rendent cette mesure d'angle aléatoire et cette variabilité semble bien modélisée par une loi normale (ce genre d'hypothèse aussi peut faire l'objet d'un test comme nous le verrons plus tard).

En piochant aléatoirement n tranches produites, nous disposons des réalisations de *v.a.i.i.d.* $X_1, \dots, X_n$ de loi normale $\mathcal{N}(m, \sigma^2)$ de moyenne $m \in \mathbb{R}$ **inconnue** et de variance $\sigma^2 > 0$ **connue** (idem, il y a des tests pour valider cela). Nous voulons ainsi tester

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

pour un niveau de confiance $1 - \alpha \in]0, 1[$.

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

Pour construire une règle de décision, nous considérons la moyenne empirique

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

qui est un estimateur sans biais et consistant de m.

Loi de $\overline{X}_n$

Une combinaison linéaire de variables normales **indépendantes** suit une loi normale. Pour la moyenne empirique, nous obtenons

$$\overline{X}_n \text{ suit la loi } \mathcal{N}\left(m, \frac{\sigma^2}{n}\right).$$

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

Guidés par l'idée d'un intervalle de confiance de niveau $1-\alpha$, nous proposons la règle de décision suivante :

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant x_\alpha$$

où $x_\alpha \in \mathbb{R}$ est tel que

$$\mathbb{P}_{H_0}(\overline{X}_n \leqslant x_\alpha) \leqslant \alpha.$$

Autrement dit, nous souhaitons rejeter l'hypothèse d'une moyenne supérieure à 42 lorsque la moyenne empirique observée sur notre échantillon est « trop petite ».

Question : comment **calibrer** x_α pour assurer la probabilité d'erreur de 1ère espèce ?

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

Une première étape consiste à introduire la version Z **centrée et réduite** de la moyenne empirique $\overline{X}_n$,

$$Z = \sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}}.$$

Ainsi, nous avons $\overline{X}_n = m + \sqrt{\dfrac{\sigma^2}{n}}Z$ avec Z qui suit une loi $\mathcal{N}(0,1)$.

De cette façon, nous avons exhibé une **loi bien connue qui ne dépend pas de** m et nous cherchons donc $x_\alpha \in \mathbb{R}$ tel que

$$\mathbb{P}_{H_0}\left(Z \leqslant \sqrt{n}\frac{x_\alpha - m}{\sqrt{\sigma^2}}\right) \leqslant \alpha.$$

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

La borne dans la probabilité dépend de m qui reste inconnue **même sous l'hypothèse** H_0. Cependant, si $m > 42$ alors nous savons que

$$\sqrt{n}\frac{x_\alpha - m}{\sqrt{\sigma^2}} < \sqrt{n}\frac{x_\alpha - 42}{\sqrt{\sigma^2}} = z_\alpha$$

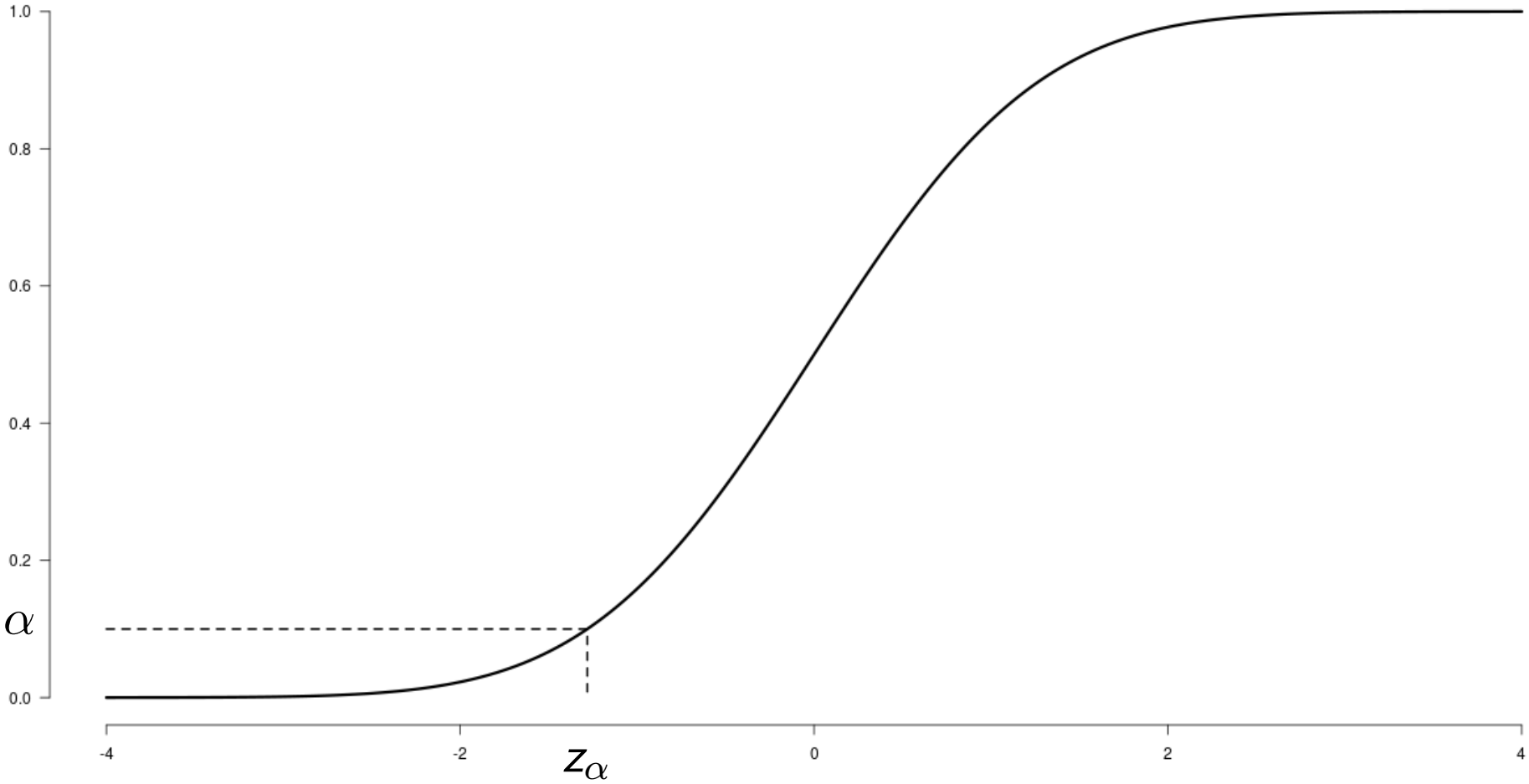
et donc

$$\mathbb{P}_{H_0}\left(Z \leqslant \sqrt{n}\frac{x_\alpha - m}{\sqrt{\sigma^2}}\right) \leqslant \mathbb{P}\left(Z \leqslant z_\alpha\right) \qquad \text{(probabilité libre de } H_0\text{)}$$

Il suffit de prendre $z_\alpha \in \mathbb{R}$ comme le **quantile** de niveau α de la loi $\mathcal{N}(0,1)$,

$$F_{\mathcal{N}(0,1)}(z_\alpha) = \int_{-\infty}^{z_\alpha} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt = \alpha \qquad \text{(Merci monsieur l'ordinateur !)}$$

Exemple du pain d'épice



Fonction de répartition de la loi $\mathcal{N}(0,1)$.

Exemple : pour $\alpha = 5\%$, nous obtenons $z_\alpha = -1.644854\ldots$

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

À partir de cette valeur de z_α, nous déduisons le seuil de rejet x_α,

$$\sqrt{n}\frac{x_\alpha - 42}{\sqrt{\sigma^2}} = z_\alpha \iff x_\alpha = 42 + z_\alpha\sqrt{\frac{\sigma^2}{n}}.$$

La règle de décision est donc donnée par

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha\sqrt{\frac{\sigma^2}{n}}.$$

Remarque : il s'agit bien d'une règle **statistique** car le seuil est **calculable** puisque tout est connu, en particulier la variance σ^2 dans cet exemple.

Exemple du pain d'épice

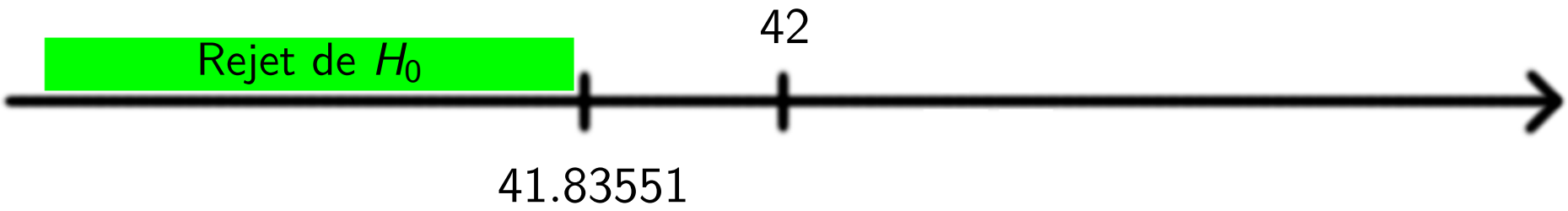
$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}}$$

Appliquons ce test avec $\alpha = 5\%$ (donc $z_\alpha = -1.644854$), $n = 100$ tranches de pain d'épice et une variance $\sigma^2 = 1$.
La règle de décision devient

$$\text{Rejeter } H_0 \iff \overline{X}_{100} \leqslant 41.83551.$$

Si la **réalisation** $\overline{x}_{100}$ de la **variable aléatoire** $\overline{X}_{100}$ sur notre échantillon donne $\overline{x}_{100} = 42.126$, alors **nous acceptons l'hypothèse** « $m > 42$ ».



Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}}$$

- **Erreur de 1ère espèce :** dans **moins de** 5% **des cas**, nous rejetons la production de pain d'épice alors qu'elle est de qualité, gaspillage !
  $\Rightarrow$ Le patron ne va pas être content (en fait, il ne le saura pas ...).
- **Erreur de 2ème espèce :** avec une **probabilité d'autant plus petite** que m est inférieure à 42 (puissance du test), nous vendons des tranches de mauvaise qualité.
  $\Rightarrow$ Le client ne va pas être content (lui, il le saura ...).

Exemple du pain d'épice

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}}$$

- **Erreur de 1ère espèce :** dans **moins de** 5% **des cas**, nous rejetons la production de pain d'épice alors qu'elle est de qualité, gaspillage !
  $\Rightarrow$ Le patron ne va pas être content (en fait, il ne le saura pas ...).
- **Erreur de 2ème espèce :** avec une **probabilité d'autant plus petite** que m est inférieure à 42 (puissance du test), nous vendons des tranches de mauvaise qualité.
  $\Rightarrow$ Le client ne va pas être content (lui, il le saura ...).

Et si nous prenions le point de vue du client ?

Exemple du pain d'épice (point de vue du client)

Pour la personne qui achète notre pain d'épice, l'important est surtout de ne pas trouver des tranches de mauvaise qualité. En termes de test statistique, cela signifie qu'elle veut s'assurer que l'erreur contrôlée est celle commise sur l'hypothèse « $m \leqslant 42$ » qui servait d'alternative pour notre test initial.

Le point de vue du client consiste donc à échanger les hypothèses précédentes et à tester

$$H_0' \ : \ m \leqslant 42 \qquad \text{contre} \qquad H_1' \ : \ m > 42$$

pour un niveau de confiance $1 - \alpha \in]0, 1[$.
Les mêmes idées que précédemment nous conduisent à proposer la règle de décision

$$\text{Rejeter } H_0' \iff \overline{X}_n > x_\alpha'$$

où $x_\alpha' \in \mathbb{R}$ est tel que

$$\mathbb{P}_{H_0'}\left(\overline{X}_n > x_\alpha'\right) \leqslant \alpha.$$

Exemple du pain d'épice (point de vue du client)

$$H_0' \ : \ m \leqslant 42 \qquad \text{contre} \qquad H_1' \ : \ m > 42$$

Afin de calibrer le seuil x'_α, nous procédons de la même façon en introduisant le nombre $z'_\alpha \in \mathbb{R}$ tel que

$$\int_{z'_\alpha}^{+\infty} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt = \alpha \iff F_{\mathcal{N}(0,1)}(z'_\alpha) = \int_{-\infty}^{z'_\alpha} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt = 1 - \alpha.$$

Sous l'hypothèse que $m \leqslant 42$, nous savons

$$\sqrt{n}\frac{x'_\alpha - m}{\sqrt{\sigma^2}} \geqslant \sqrt{n}\frac{x'_\alpha - 42}{\sqrt{\sigma^2}}$$

et cela conduit à

$$x'_\alpha = 42 + z'_\alpha \sqrt{\frac{\sigma^2}{n}}.$$

Exemple du pain d'épice (point de vue du client)

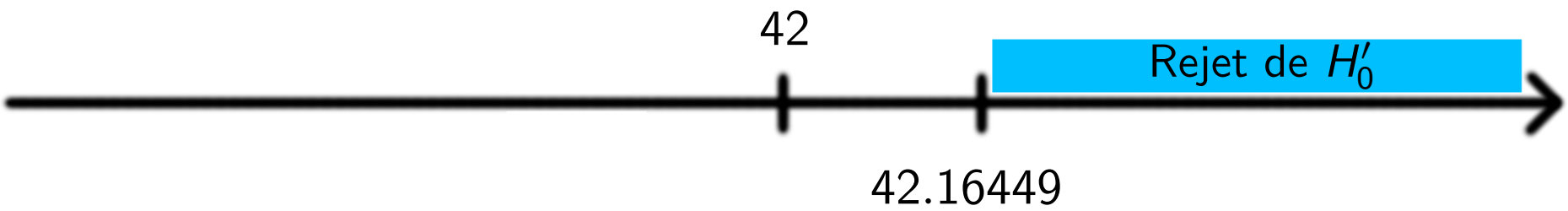
$$H_0' \ : \ m \leqslant 42 \qquad \text{contre} \qquad H_1' \ : \ m > 42$$

$$\text{Rejeter } H_0' \iff \overline{X}_n > 42 + z'_\alpha \sqrt{\frac{\sigma^2}{n}}$$

Appliquons ce test avec les mêmes valeurs que dans l'exemple précédent : $\alpha = 5\%$ (et $z'_\alpha = 1.644854$), $n = 100$ et $\sigma^2 = 1$.
La règle de décision devient

$$\text{Rejeter } H_0' \iff \overline{X}_n > 42.16449.$$

Si la **moyenne observée** vaut $\overline{x}_{100} = 42.126$, alors **nous acceptons l'hypothèse** « $m \leqslant 42$ ».



Exemple du pain d'épice (point de vue du client)

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}} \ (= 41.83551)$$

$$H_0' \ : \ m \leqslant 42 \qquad \text{contre} \qquad H_1' \ : \ m > 42$$

$$\text{Rejeter } H_0' \iff \overline{X}_n > 42 + z_\alpha' \sqrt{\frac{\sigma^2}{n}} \ (= 42.16449)$$

Nous venons de construire deux tests de même niveau de confiance qui donnent des **réponses opposées** sur les **mêmes données** . . .

Est-ce contradictoire ?

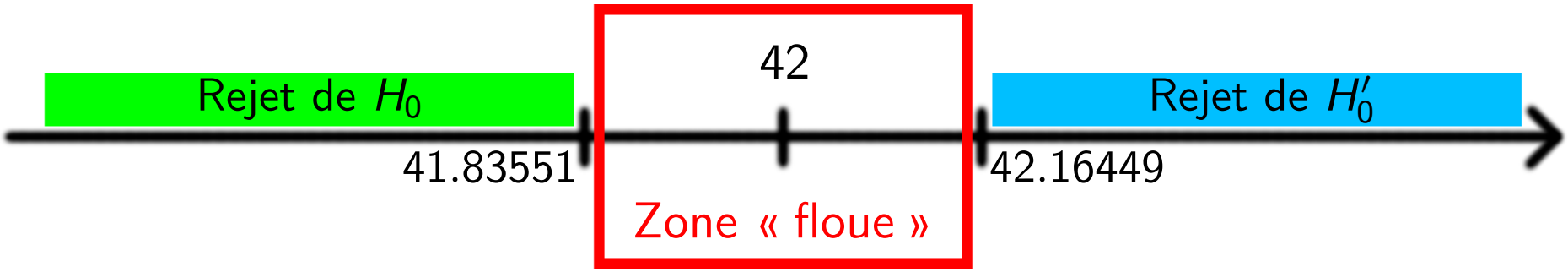
Exemple du pain d'épice (point de vue du client)

$$H_0 \ : \ m > 42 \qquad \text{contre} \qquad H_1 \ : \ m \leqslant 42$$

$$\text{Rejeter } H_0 \iff \overline{X}_n \leqslant 42 + z_\alpha \sqrt{\frac{\sigma^2}{n}} \ (= 41.83551)$$

$$H_0' \ : \ m \leqslant 42 \qquad \text{contre} \qquad H_1' \ : \ m > 42$$

$$\text{Rejeter } H_0' \iff \overline{X}_n > 42 + z_\alpha' \sqrt{\frac{\sigma^2}{n}} \ (= 42.16449)$$



Choix de l'hypothèse nulle H_0

Dans le « doute », un test statistique favorise **toujours** son hypothèse nulle. Les hypothèses ne sont pas **symétriques** et **la décision dépend du parti pris de départ**.

Idée générale : un test statistique ne rejette son hypothèse nulle que si celle-ci n'est vraiment **pas vraisemblable**. Nous parlons alors de **significativité** du test statistique.

Choix de l'hypothèse nulle H_0

Dans le « doute », un test statistique favorise **toujours** son hypothèse nulle. Les hypothèses ne sont pas **symétriques** et **la décision dépend du parti pris de départ**.

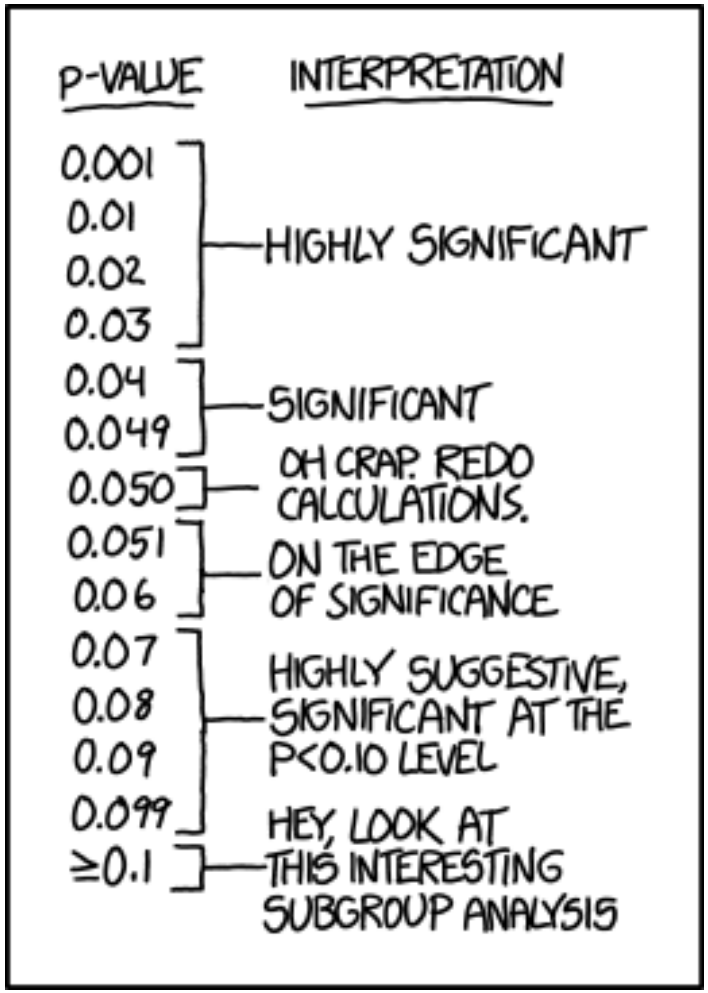
Idée générale : un test statistique ne rejette son hypothèse nulle que si celle-ci n'est vraiment **pas vraisemblable**. Nous parlons alors de **significativité** du test statistique.

Comment choisir H_0 ?

1. H_0 est l'hypothèse la plus grave (le pont s'écroule, le médicament est mortel, ...).
2. H_0 est l'hypothèse la plus communément admise.
3. Les calculs de calibration ne peuvent être faits que sous H_0.

Notion de p-valeur

L'utilisation de la p-valeur est très criticable en pratique.



P-Values, XKCD, xkcd.com/1478

« We teach it because it's what we do ; we do it because it's what we teach. »

George Cobb

Notion de p-valeur

Bien que la définition de la p-valeur fasse **encore débat**, il s'agit d'une quantité couramment utilisée dans de nombreux domaines de recherche.

L'objectif de la p-valeur est de quantifier le « **degré de significativité** » d'un test statistique.

Une façon commune de présenter la p-valeur est de la définir comme la **plus petite valeur** de l'erreur de 1ère espèce $\alpha \in]0, 1[$ pour laquelle les observations conduisent au **rejet de l'hypothèse nulle** H_0.

La p-valeur est donc la probabilité, sous l'hypothèse H_0, d'observer les données les « plus extrêmes ».

Lien entre le niveau et la p-valeur

$$\text{Rejet de } H_0 \text{ au niveau } 1-\alpha \iff p\text{-valeur} < \alpha.$$

Plus la p-valeur est faible, plus le risque de rejeter H_0 à tort est faible.

Bootstrap : motivation

On observe un échantillon

$$X_1, \ldots, X_n \sim P,$$

et on calcule un estimateur

$$\widehat{\theta}_n = T(X_1, \ldots, X_n).$$

Comme tout estimateur, $\widehat{\theta}_n$ est une variable aléatoire.

Problème

Pour construire un intervalle de confiance, il faudrait connaître ou approximer la loi de

$$\widehat{\theta}_n - \theta.$$

Le TCL donne une réponse pour les moyennes, mais pas directement pour une médiane, un quantile, une erreur de prédiction, etc.

Bootstrap non paramétrique

Essayer de remplacer la loi inconnue P par la loi empirique

$$P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}.$$

Conditionnellement aux données, on tire

$$X_1^*, \dots, X_n^* \sim P_n,$$

c'est-à-dire que l'on tire avec remise dans l'échantillon observé.

L'idée de remplacer P par P_n vient d'un théorème affirmant que, sous certaines hypothèses, on a

$$P_n \underset{n\to\infty}{\longrightarrow} P.$$

Au sens de la convergence en faible.

Bootstrap non paramétrique

Algorithme

1. Calculer $\widehat{\theta}_n = T(X_1, \ldots, X_n)$.
2. Pour $b = 1, \ldots, B$, tirer un échantillon bootstrap $(X_1^{*(b)}, \ldots, X_n^{*(b)})$.
3. Calculer $\widehat{\theta}_n^{*(b)} = T(X_1^{*(b)}, \ldots, X_n^{*(b)})$.
4. Utiliser la dispersion des $\widehat{\theta}_n^{*(b)}$ pour approximer l'incertitude de $\widehat{\theta}_n$.

Théorème bootstrap

Théorème

Soient $X_1, \ldots, X_n$ des variables i.i.d. de loi P, d'espérance m et, pour $b = 1, \ldots, B$, un échantillon bootstrap $(X_1^{*(b)}, \ldots, X_n^{*(b)})$.
Si il existe une variable aléatoire Z telle que

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\mathcal{L}} Z,$$

alors sous des hypothèses techniques conditionnellement à $X_1, \ldots, X_n$ on a

$$\sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_n^{*(b)} - \hat{\theta}_n\right) \xrightarrow[B\to\infty]{\mathcal{L}} Z$$

En conséquence, on est capable d'échantillonner approximativement Z en échantillonnant $\sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_n^{*(b)} - \hat{\theta}_n\right)$

Bootstrap : intervalles de confiance

Pour construire une intervalle de confiance de niveau $\alpha \in (0,1)$ pour θ (pour l'estimation ou le test) on a besoin, par exemple, de trouver $z_{\alpha/2}$ et $z_{1-\alpha/2}$ tels que

$$\mathbb{P}(z_{\alpha/2} \leq Z \leq z_{1-\alpha/2}) = 1 - \alpha.$$

Problème et solution

Si Z est inconnu alors $z_{\alpha/2}$ et $z_{1-\alpha/2}$ le sont aussi.
On va alors les estimer en échantillonnant Z avec $\sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_n^{*(b)} - \hat{\theta}_n\right)$.

Bootstrap : procédure Monte Carlo

On se fixe un nombre de répétition K et pour chaque $1 \leq k \leq K$,

1. On choisit un nombre de réplication bootstrap B
2. Pour $b = 1, \ldots, B$, on tire un échantillon bootstrap $(X_1^{*(b)}, \ldots, X_n^{*(b)})$.
3. On construit $\tilde{Z}_k := \sqrt{B}\left(\frac{1}{B}\sum_{b=1}^{B} \hat{\theta}_n^{*(b)} - \hat{\theta}_n\right)$.

Ainsi l'échantillon $(\tilde{Z}_1, \ldots, \tilde{Z}_K)$ est tiré approximativement suivant la loi de Z.
On peut estimer $z_{\alpha/2}$ et $z_{1-\alpha/2}$ à l'aide des quantiles empiriques des $(\tilde{Z}_1, \ldots, \tilde{Z}_K)$.

Tests multiples : motivation

Un seul test

Pour un test de niveau α, on contrôle l'erreur de première espèce :

$$\mathbb{P}_{H_0}(\text{rejeter } H_0) \leq \alpha.$$

Beaucoup de tests

Si l'on teste m hypothèses vraies au niveau α, alors, sous indépendance,

$$\mathbb{P}(\text{au moins un faux positif}) = 1 - (1-\alpha)^m.$$

Par exemple, avec $m = 100$ et $\alpha = 5\%$,

$$1 - 0.95^{100} \simeq 99.4\%.$$

Correction de Bonferroni

Définition : FWER

Pour m tests, le *Family-Wise Error Rate* est

$$\mathrm{FWER} = \mathbb{P}(V \geq 1),$$

où V est le nombre de fausses découvertes, c'est-à-dire le nombre d'hypothèses nulles vraies qui sont rejetées.

Théorème de Bonferroni

Si chaque hypothèse $H_{0,j}$ est testée au niveau α_j, alors

$$\mathrm{FWER} \leq \sum_{j \in \mathcal{H}_0} \alpha_j \leq \sum_{j=1}^{m} \alpha_j,$$

où $\mathcal{H}_0$ désigne l'ensemble des hypothèses nulles vraies. En particulier, si $\alpha_j = \alpha/m$, alors

FDR et procédure de Benjamini–Hochberg

Définition : FDR (false discovery rate)

Si R est le nombre total de rejets et V le nombre de faux rejets, alors

$$\mathrm{FDR} = \mathbb{E}\left[\frac{V}{\max(R,1)}\right].$$

Le FDR contrôle la proportion moyenne de fausses découvertes parmi les découvertes annoncées.

Procédure BH au niveau q

On ordonne les p-valeurs :

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

On pose

$$\widehat{k} = \max \left\{ k : p_{(k)} \leq \frac{kq}{m} \right\},$$

puis on rejette les hypothèses associées à $p_{(1)}, \ldots, p_{(\widehat{k})}$.

Théorème de Benjamini–Hochberg

Sous indépendance des p-valeurs et lorsque les p-valeurs correspondant aux vraies hypothèses nulles sont uniformes sur $[0, 1]$, la procédure BH vérifie

$$\mathrm{FDR} \leq \frac{m_0}{m} q \leq q,$$

où m_0 est le nombre d'hypothèses nulles vraies.

Question statistique $\leftrightarrow$ test à utiliser

Question	**Exemple de test**
Une moyenne est-elle égale à une valeur de référence ?	Student à un échantillon test asymptotique par TCL
Deux groupes ont-ils la même moyenne ?	Test de Student Test de Welch
Plusieurs groupes ont-ils la même moyenne ?	ANOVA
Deux proportions sont-elles différentes ?	Test binomial Test du χ^2 Test asymptotique par TCL
Deux variables qualitatives sont-elles indépendantes ?	Test du χ^2 d'indépendance Test exact de Fisher
Une variable suit-elle une loi donnée ? Les données sont-elles normales ?	Test de Kolmogorov–Smirnov Shapiro–Wilk

ATTENTION !

Avant d'utiliser un test vérifiez que les données satisfont les hypothèses du test.... ou testez les.

5.0 Quelques tests statistiques classiques

Tests sur la moyenne (loi normale)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Exemples d'hypothèses :

$$
\begin{array}{lllll}
H_0 : m = m_0 & \text{contre} & H_1 : m = m_1 \ (\text{avec } m_0 \neq m_1) \\
H_0 : m = m_0 & \text{contre} & H_1 : m > m_0 \ (\text{ou } m < m_0) \\
H_0 : m = m_0 & \text{contre} & H_1 : m \neq m_0 \\
H_0 : m \leqslant m_0 & \text{contre} & H_1 : m > m_0 \\
H_0 : m \geqslant m_0 & \text{contre} & H_1 : m < m_0
\end{array}
$$

Et bien d'autres ...

Tests sur la moyenne (loi normale)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Si la variance σ^2 est **connue**, la règle décision se construit à partir de la moyenne empirique $\overline{X}_n$ et se calibre avec

$$\sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\sigma^2}} \text{ suit la loi } \mathcal{N}(0,1).$$

Voir l'exemple du pain d'épice ...

Tests sur la moyenne (loi normale)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Si la variance σ^2 est **inconnue**, elle peut être estimée par

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

Cette définition fait intervenir la somme des carrés de variables normales indépendantes recentrées par la moyenne empirique. À la variance σ^2 près, une telle variable admet une loi dite du χ^2 à $n-1$ degrés de liberté (et non pas n à cause du **recentrage empirique**),

$$\frac{1}{\sigma^2} \sum_{k=1}^{n} \left(X_k - \overline{X}_n \right)^2 \text{ suit la loi } \chi^2(n-1).$$

Conséquence : $\mathbb{E}[\hat{\sigma}_n^2] = \dfrac{n-1}{n}\sigma^2$ donc $b(\hat{\sigma}_n^2) = -\sigma^2/n \neq 0$.

Tests sur la moyenne (loi normale)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Un estimateur **sans biais** de la variance σ^2 est donné par

$$\tilde{\sigma}_n^2 = \frac{n}{n-1}\hat{\sigma^2}_n = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_k - \overline{X}_n\right)^2.$$

Le théorème de Cochran assure que la moyenne empirique $\overline{X}_n$ est **indépendante** de $\tilde{\sigma}_n^2$ (et de $\hat{\sigma}_n^2$). Cela permet de déduire que

$$\sqrt{n}\frac{\overline{X}_n - m}{\sqrt{\tilde{\sigma}_n^2}}$$

suit la loi de Student $\mathcal{T}(n-1)$ à $n-1$ degrés de liberté. Cette loi permet de calibrer la règle de décision de façon similaire au cas de variance connue.

Tests sur la moyenne (cas général)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* avec $\mathbb{E}[X_1] = m \in \mathbb{R}$ et $\mathrm{Var}(X_1) = \sigma^2 > 0$.

Exemples d'hypothèses :

$$
\begin{array}{lllll}
H_0 & : & m = m_0 & \text{contre} & H_1 : m = m_1 \ (\text{avec } m_0 \neq m_1) \\
H_0 & : & m = m_0 & \text{contre} & H_1 : m > m_0 \ (\text{ou } m < m_0) \\
H_0 & : & m = m_0 & \text{contre} & H_1 : m \neq m_0 \\
H_0 & : & m \leqslant m_0 & \text{contre} & H_1 : m > m_0 \\
H_0 & : & m \geqslant m_0 & \text{contre} & H_1 : m < m_0
\end{array}
$$

Et bien d'autres ...

Remarque : bien qu'il soit possible dans certains cas de proposer des tests de niveau fixé, les résultats généraux sont tous **asymptotiques**.

Tests sur la moyenne (cas général)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* avec $\mathbb{E}[X_1] = m \in \mathbb{R}$ et $\mathrm{Var}(X_1) = \sigma^2 > 0$.

Si la variance σ^2 est **connue**, la règle décision se construit à partir de la moyenne empirique $\overline{X}_n$ et se **calibre asymptotiquement** comme dans le cas normal grâce au théorème central limite,

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

Si la variance σ^2 est **inconnue**, cette convergence en loi reste vrai en remplaçant la variance par $\hat{\sigma}_n^2$ (ou $\tilde{\sigma}_n^2$) car il s'agit d'un estimateur consistant et le lemme de Slutsky donne

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

Cela permet encore de **calibrer asymptotiquement** la règle de décision.

Tests sur la variance (loi normale)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Exemples d'hypothèses :

$$
\begin{array}{lllll}
H_0 : \sigma^2 = \sigma_0^2 & \text{contre} & H_1 : \sigma^2 = \sigma_1^2 \ (\text{avec } \sigma_0^2 \neq \sigma_1^2) \\
H_0 : \sigma^2 = \sigma_0^2 & \text{contre} & H_1 : \sigma^2 > \sigma_0^2 \ (\text{ou } \sigma^2 < \sigma_0^2) \\
H_0 : \sigma^2 = \sigma_0^2 & \text{contre} & H_1 : \sigma^2 \neq \sigma_0^2 \\
H_0 : \sigma^2 \leqslant \sigma_0^2 & \text{contre} & H_1 : \sigma^2 > \sigma_0^2 \\
H_0 : \sigma^2 \geqslant \sigma_0^2 & \text{contre} & H_1 : \sigma^2 < \sigma_0^2
\end{array}
$$

Et bien d'autres ...

Tests sur la variance (loi normale)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{N}(m, \sigma^2)$

Si la moyenne m est **connue**, il est possible de recentrer les variables observées de façon **déterministe** et de calibrer la règle de décision avec

$$\frac{1}{\sigma^2} \sum_{k=1}^{n} (X_k - m)^2 \text{ suit la loi } \chi^2(n).$$

Si la moyenne m est **inconnue**, il faut **recentrer empiriquement** avec $\overline{X}_n$ et la calibration se fait grâce à

$$\frac{1}{\sigma^2} \sum_{k=1}^{n} \left(X_k - \overline{X}_n\right)^2 \text{ suit la loi } \chi^2(n-1).$$

Test d'égalité des variances (loi normale)

Cadre : deux groupes **indépendants** de variables $X_1, \dots, X_p$ *i.i.d.* de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \dots, Y_q$ *i.i.d.* de loi $\mathcal{N}(m_2, \tau^2)$

$$H_0 \ : \ \sigma^2 = \tau^2 \qquad \text{contre} \qquad H_1 \ : \ \sigma^2 \neq \tau^2$$

Nous considérons les estimateurs sans biais des variances,

$$\tilde{\sigma}_p^2 = \frac{1}{p-1} \sum_{k=1}^{p} \left(X_k - \overline{X}_p \right)^2 \qquad \text{et} \qquad \tilde{\tau}_q^2 = \frac{1}{q-1} \sum_{k=1}^{q} \left(Y_k - \overline{Y}_q \right)^2 .$$

Sous l'hypothèse H_0 d'égalité des variances, nous avons

$$F = \frac{\tilde{\sigma}_p^2}{\tilde{\tau}_q^2} = \frac{\tilde{\sigma}_p^2/\sigma^2}{\tilde{\tau}_q^2/\tau^2} \quad \text{suit la loi} \quad \frac{\chi^2(p-1)/(p-1)}{\chi^2(q-1)/(q-1)}.$$

Il s'agit de la loi de Fisher $\mathcal{F}(p-1, q-1)$ à $p-1$ et $q-1$ degrés de liberté.

Test d'égalité des variances (loi normale)

Cadre : deux groupes **indépendants** de variables $X_1, \dots, X_p$ *i.i.d.* de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \dots, Y_q$ *i.i.d.* de loi $\mathcal{N}(m_2, \tau^2)$

$$H_0 \ : \ \sigma^2 = \tau^2 \qquad \text{contre} \qquad H_1 \ : \ \sigma^2 \neq \tau^2$$

Le principe de la règle de décision est de rejeter H_0 si le rapport $F = \tilde{\sigma}_p^2 / \tilde{\tau}_q^2$ est « trop petit » ou « trop grand »,

$$\text{Rejeter } H_0 \iff F < u_\alpha \text{ ou } F > v_\alpha$$

avec u_α et v_α à calibrer pour un niveau $1 - \alpha \in]0, 1[$ donné.

Remarques : pour calibrer u_α et v_α, il faut utiliser le fait que, sous H_0,

$$F \text{ suit la loi } \mathcal{F}(p-1, q-1) \quad \text{et} \quad 1/F \text{ suit la loi } \mathcal{F}(q-1, p-1).$$

Test de comparaison des moyennes (loi normale)

Cadre : deux groupes **indépendants** de variables $X_1, \dots, X_p$ *i.i.d.* de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \dots, Y_q$ *i.i.d.* de loi $\mathcal{N}(m_2, \sigma^2)$ de **même variance**

Exemples d'hypothèses :

$$
\begin{array}{lcll}
H_0 : m_1 = m_2 & \text{contre} & H_1 : m_1 \neq m_2 \\
H_0 : m_1 = m_2 & \text{contre} & H_1 : m_1 > m_2 \\
H_0 : m_1 = m_2 & \text{contre} & H_1 : m_1 < m_2 \\
H_0 : m_1 \leqslant m_2 & \text{contre} & H_1 : m_1 > m_2 \\
H_0 : m_1 \geqslant m_2 & \text{contre} & H_1 : m_1 < m_2
\end{array}
$$

Et bien d'autres ...

Pour estimer la variance σ^2 commune sans biais, nous disposons de

$$\tilde{\sigma}^2_{p,q} = \frac{(p-1)\tilde{\sigma}^2_{X,p} + (q-1)\tilde{\sigma}^2_{Y,q}}{p+q-2}.$$

Test de comparaison des moyennes (loi normale)

Cadre : deux groupes **indépendants** de variables $X_1, \dots, X_p$ *i.i.d.* de loi $\mathcal{N}(m_1, \sigma^2)$ et $Y_1, \dots, Y_q$ *i.i.d.* de loi $\mathcal{N}(m_2, \sigma^2)$ de **même variance**

Le théorème de Cochran donne encore que $\overline{X}_p$ et $\overline{Y}_q$ sont indépendantes de $\tilde{\sigma}^2_{p,q}$ et que

$$(p+q-2)\frac{\tilde{\sigma}^2_{p,q}}{\sigma^2} \text{ suit la loi } \chi^2(p+q-2).$$

La règle de décision se calibre alors grâce à

$$\sqrt{\frac{pq}{p+q}} \times \frac{(\overline{X}_p - m_1) - (\overline{Y}_q - m_2)}{\sqrt{\tilde{\sigma}^2_{p,q}}} \text{ suit la loi } \mathcal{T}(p+q-2).$$

Test de Shapiro-Wilk (non paramétrique)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{L}$ **inconnue**

Le test de normalité de Shapiro-Wilk considère

$$H_0 \ : \ \mathcal{L} \text{ est normale} \qquad \text{contre} \qquad H_1 \ : \ \mathcal{L} \text{ n'est pas normale.}$$

Pour cela, nous introduisons la **version ordonnée** des variables observées,

$$X_{(1)} \leqslant \cdots \leqslant X_{(n)}$$

et nous définissons la variable

$$W = \frac{\left(\sum_{k=1}^{n} a_k X_{(k)} \right)^2}{\sum_{k=1}^{n} (X_k - \overline{X}_n)^2}.$$

Test de Shapiro-Wilk (non paramétrique)

Les coefficients $a_1, \ldots, a_n$ sont connus et disponibles dans tout bon logiciel de statistique. Ils sont donnés par

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} = \frac{m^\top \Sigma^{-1}}{\sqrt{m^\top \Sigma^{-2} m}}$$

où $m \in \mathbb{R}^n$ est le vecteur des espérances de la version ordonnée de n *v.a.i.i.d.* normales centrées réduites et Σ est la matrice de covariance de ces mêmes variables normales ordonnées.

En pratique, plus la valeur de W est élevée, plus l'adéquation à la loi normale est acceptable,

$$\text{Rejeter } H_0 \iff W < w_\alpha.$$

Test de significativité en régression

On cherche à vérifier si une régression linéaire est pertinente, i.e. si le modèle

$$Y_i \sim \mathcal{N}(ax_i + b, \sigma^2) \quad i = 1, \dots, n$$

est bien ajusté. Ce qui revient formellement à tester

$$H_0 : a = 0 \quad VS \quad H_1 : a \neq 0.$$

On considère la statistique

$$F = (n-2)\frac{R^2}{1-R^2} \sim \mathcal{F}(1, n-2).$$

Ce qui donne la règle de décision suivante :

$$\text{Rejeter } H_0 \iff F < f_\alpha.$$

Test de Kolmogorov-Smirnov (non paramétrique)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{L}_X$ **inconnue**

Pour une loi $\mathcal{L}$ donnée, le test d'adéquation de Kolmogorov-Smirnov considère

$$H_0 \ : \ \mathcal{L}_X = \mathcal{L} \qquad \text{contre} \qquad H_1 \ : \ \mathcal{L}_X \neq \mathcal{L}.$$

Nous introduisons la **fonction de répartition empirique** des observations,

$$\forall x \in \mathbb{R}, \ F_n(x) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{X_k \leqslant t}.$$

Test de Kolmogorov-Smirnov (non paramétrique)

Cadre : $X_1, \dots, X_n$ *v.a.i.i.d.* de loi $\mathcal{L}_X$ **inconnue**

Pour une loi $\mathcal{L}$ donnée, le test d'adéquation de Kolmogorov-Smirnov considère

$$H_0 \ : \ \mathcal{L}_X = \mathcal{L} \qquad \text{contre} \qquad H_1 \ : \ \mathcal{L}_X \neq \mathcal{L}.$$

Si F est la fonction de répartition de $\mathcal{L}$, il est possible de montrer que, sous l'hypothèse H_0, la « **fonction aléatoire** »

$$x \in \mathbb{R} \mapsto \sqrt{n}(F_n(x) - F(x))$$

converge en loi vers un **pont brownien**. Cet objet aléatoire dépasse de loin le cadre de ce cours mais il est bien connu et permet de calibrer **asymptotiquement** la règle de décision suivante

$$\text{Rejeter } H_0 \iff \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > k_\alpha.$$

Test d'interdependence (cas Gaussien)

Cadre : $(Y_1, X_1), \dots, (X_n, Y_n)$ des vecteur gaussiens *i.i.d.*, i.e

$$(X_i, Y_i) \sim \mathcal{N}_2 \left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \right).$$

avec ρ le coefficient de correlation.

Dans le des vecteurs gaussiens, on sait que $X \perp Y \iff \rho = 0$. On considère alors le test

$$H_0 : \rho = 0 \quad \text{contre} \quad H_1 : \rho \neq 0.$$

Test d'interdependence (cas Gaussien)

Le coefficient de corrélation empirique est :

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

Il est possible de montrer que sous $H_0 : \rho = 0$, la statistique

$$\frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \sim \mathcal{T}(n-2).$$

Ce qui donne la régle de décision suivante :

$$\text{Accepter } H_0 \iff t_{1-\alpha/2}(n-2) \leq \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \leq t_{\alpha/2}(n-2)$$

Analyse de la variance (ANOVA)

Cadre : $(X_{i,k})$ indépendants avec $k = 1, \dots, K$ et $i = 1, \dots n_k$ modélisé par

$$X_{i,k} = \mu + \alpha_k + \mathcal{N}(0, \sigma^2).$$

Avec $\mu \in \mathbb{R}$ la moyenne globale, les α_k un effet du groupe k.
on veut tester

$$H_0 : \alpha_1 = \cdots = \alpha_K = 0 \quad \text{contre} \quad H_1 : \exists\ (k, l) \text{ tels que } k \neq l,\ \alpha_k \neq \alpha_l.$$

Remarques :

- Cela signifie qu'on cherche à tester si les K groupes ont la **même moyenne**.
- C'est beaucoup **plus puissant** que de tester $\alpha_l = \alpha_k$ pour toutes les paires (k, l).

Analyse de la variance (ANOVA)

On se sert de la décomposition de la variance (classique dans les modèles linéaires) pour construire la statistique de test : SCT = SCE + SCR, avec

- Somme des carrés totale (variance totale) :

$$\text{SCT} = \sum_{k=1}^{K}\sum_{k=i}^{n_k}(X_{i,k} - \bar{X}_n)^2, \quad \text{avec } \bar{X}_n = \sum_{k=1}^{K}\sum_{k=i}^{n_k} X_{i,k}.$$

- Somme des carrées expliquée (variance interigroupes) :

$$\text{SCE} = \sum_{k=1}^{K} n_k(\bar{X}_{(k)} - \bar{Y}_n)^2 \quad \text{avec } \bar{X}_{(k)} = \sum_{k=i}^{n_k} X_{i,k}, \text{ pour } k = 1, \ldots, K.$$

- Somme des carrés résiduels (variance intra-groupes) :

$$\text{SCR} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(X_{i,k} - \bar{X}_k)^2.$$

Analyse de la variance (ANOVA)

On a alors, sous H_0, en notant $N = \sum_{k=1}^{K} n_k$:

$$\frac{\text{SCE}}{\sigma^2} \sim \chi^2_{K-1} \quad \text{et} \quad \frac{\text{SCR}}{\sigma^2} \sim \chi^2_{N-K}, \text{ pour tout } k = 1, \dots, K.$$

On en déduit alors la statistique de test :

$$F = \frac{\text{SCE}}{K-1} \frac{N-K}{\text{SCR}} \sim \mathcal{F}_{K-1,\, N-K}.$$

Ce qui donne la régle de décision suivante :

$$\text{Accepter } H_0 \iff F < f_\alpha.$$