Probabilités et statistiques

III – Inférence et décision

G. Chênevert

27 novembre 2023

JUNIA ISEN

Au menu aujourd'hui

Théorèmes limites

Estimation paramétrique

Tests d'hypothèse

Des vecteurs aux suites aléatoires

La dernière fois : cas des vecteurs aléatoires (X, Y) de dimension 2.

Toute cette discussion s'étend « sans trop de mal » au cas général

$$\mathbf{X} = (X_1, \ldots, X_n)$$

de la dimension n.

Aujourd'hui : **suites** aléatoires

$$\mathbf{X} = (X_n)_{n=1}^{\infty} = (X_1, \ldots, X_n, \ldots)$$

et notamment notion de limite

$$\lim_{n \to \infty} X_n.$$

Limites de variables aléatoires

Cas le plus simple : suite de v.a. « constante. »

On s'intéresse donc à une suite de variables aléatoires

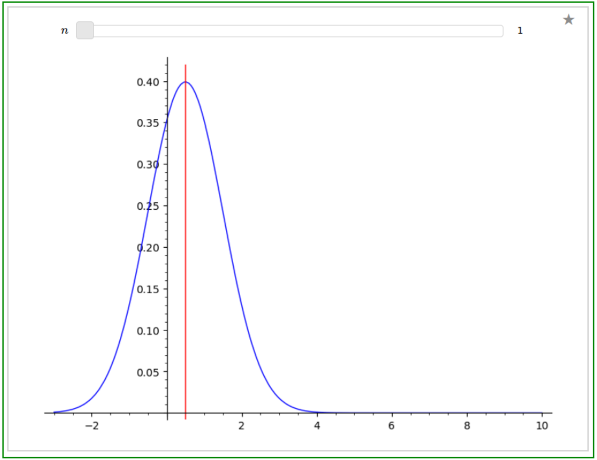
$$X_1, X_2, \dots, X_n, \dots$$

indépendantes, identiquement distribuées (i.i.d.)

Disons : espérance μ, écart-type σ

Considérons tout d'abord leur somme $T_n = X_1 + \cdots + X_n$.

$T_n = \sum_{i=1}^{n} X_i$ **avec** $X_i \sim \mathcal{N}(\frac{1}{2}, 1)$



Help | Powered by SageMath

En général

Avec $T_n = \sum_{i=1}^{n} X_i = X_1 + \cdots + X_n$:

$$\mathbb{E}[T_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n\mu$$

$$\mathrm{Var}(T_n) \overset{\text{ind}}{=} \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) = n\,\sigma^2$$

$$\Longrightarrow \sigma_{T_n} = \sqrt{n}\,\sigma$$

Moyenne échantillonnale

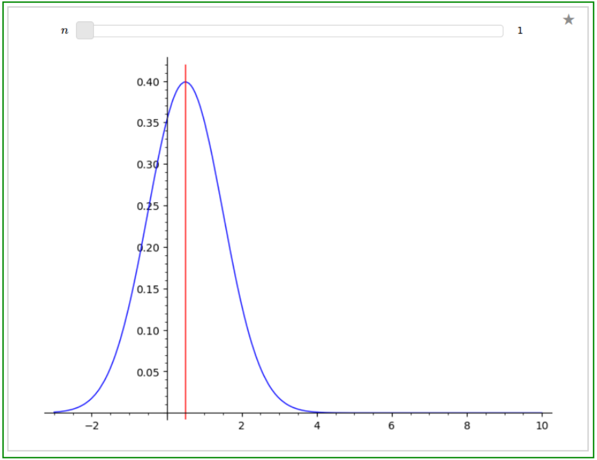
Divisons par n et formons

$$\overline{X}_n := \frac{1}{n}\,T_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \cdots + X_n}{n}.$$

Alors :

$$\mathbb{E}\left[\overline{X}_n\right] = \mu, \qquad \sigma_{\overline{X}_n} = \frac{1}{n}\,\sigma_{T_n} = \frac{\sigma}{\sqrt{n}}.$$

$\overline{X}_n$ avec $X_i \sim \mathcal{N}(\frac{1}{2}, 1)$



Help | Powered by SageMath

Loi (faible) des grands nombres

Théorème

Pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[\,|\overline{X}_n - \mu| \geq \varepsilon\,] = 0.$$

i.e. $\overline{X}_n$ *converge en probabilité* vers μ

Preuve : Inégalité de Bienaymé-Tchebychev appliquée à $\overline{X}_n$

On peut dire plus !

Écrivons

$$\overline{X}_n = \mu + \sum_{i=1}^{n} \underbrace{\frac{X_i - \mu}{n}}_{\text{esp. } 0, \text{ var. } \frac{\sigma^2}{n^2}}$$

$$\begin{aligned} \implies g_{\overline{X}_n}(t) &= e^{\mu t} \cdot \left(1 + \frac{\sigma^2}{2n^2} t^2 + \dots\right)^n \\ &= e^{\mu t} \cdot \left(1 + \frac{\sigma^2}{2n} t^2 + \dots\right) \\ &\longrightarrow e^{\mu t} \qquad \text{quand} \quad n \to \infty \end{aligned}$$

Loi (forte) des grands nombres

D'où :

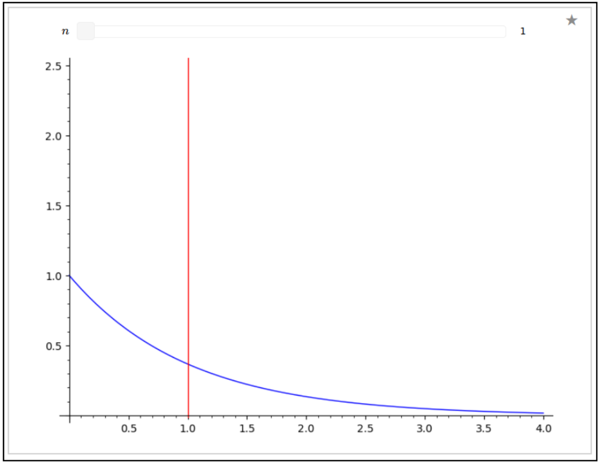
Théorème

$$\lim_{n\to\infty} \overline{X}_n = \mu \qquad \textit{presque sûrement}$$

Ou encore : $\overline{X}_n$ *converge en loi* vers une v.a. « presque constante »

(densité $\delta(x-\mu)$).

$\overline{X}_n$ pour $X_i \sim \mathcal{E}(1)$



Help | Powered by SageMath

On peut dire plus ![2]

Théorème (théorème central limite, Laplace 1809)

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \textit{ converge en loi vers une } \mathcal{N}(0,1) \textit{ quand } n \to \infty$$

En d'autres termes, pour n grand, $\overline{X}_n$ suit approximativement une

$$\mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Preuve du TCL

Si on pose $Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$, on peut écrire $Z_n = \sum_{i=1}^{n} Y_i$ avec

$$Y_i = \frac{1}{\sqrt{n}} \frac{X_i - \mu}{\sigma} \quad \text{espérance } 0\text{, variance } \frac{1}{n}$$

$$\implies g_{Y_i}(t) = 1 + \frac{t^2}{2n} + \dots$$

$$\implies g_{Z_n}(t) = \left(1 + \frac{t^2}{2n} + \dots\right)^n$$

$$\implies \ln g_{Z_n}(t) = n \ln\left(1 + \frac{t^2}{2n} + \dots\right) \sim n \cdot \left(\frac{t^2}{2n} + \dots\right) \longrightarrow \frac{t^2}{2}$$

$$\implies g_{Z_n}(t) \underset{n\to\infty}{\longrightarrow} e^{\frac{t^2}{2}} = g_Z(t) \qquad \text{avec} \qquad Z \sim \mathcal{N}(0,1)$$

Exemple : approximation normale de la binomiale

Pour les $X_i \overset{\text{iid}}{\sim} \mathcal{B}(p)$:

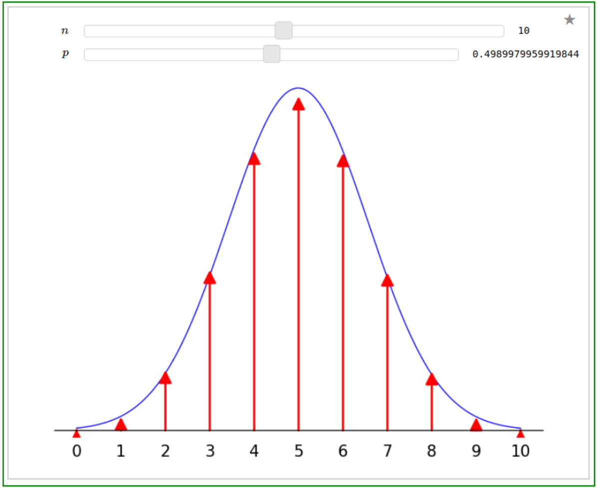
$$\overline{X}_n \rightsquigarrow \mathcal{N}\left(p, \sqrt{\frac{pq}{n}}\right) \quad \text{avec} \quad q = 1 - p.$$

Donc $\sum_{i=1}^{n} X_i = n\overline{X}_n$, de loi $\mathcal{B}(n, p)$, est approximativement

$$\mathcal{N}\left(np, \sqrt{npq}\right)$$

En pratique, approximation satisfaisante dès que np et $nq \geq 10$.

$\mathcal{B}(n, p)$ vs $\mathcal{N}\left(np, \sqrt{npq}\right)$



Help | Powered by SageMath

Au menu aujourd'hui

Théorèmes limites

Estimation paramétrique

Tests d'hypothèse

Estimation paramétrique

Une fois (sup)posé le type de modèle (loi) pour une variable qui nous intéresse, reste à déterminer « expérimentalement » les valeurs des paramètres qui y figurent

Exemples :

- p pour une $\mathcal{B}(p)$
- μ et σ pour une $\mathcal{N}(\mu, \sigma)$
- λ pour une $\mathcal{E}(\lambda)$
- a et b pour une $\mathcal{U}([a, b])$
- $\ldots$

Exemple : dé croche

Soit p la probabilité d'obtenir un 6 sur le dé ci-dessous.

Pour l'*estimer*, les ISEN62 ont gracieusement tiré un n-échantillon

$$(Y_1, Y_2, \ldots, Y_n)$$

avec $n \approx 200$ et $Y_i \in [\![1, 6]\!]$ i.i.d.

Si on note X_i la v.a. de Bernoulli associée à l'événement $Y_i = 6$, alors $X_i \sim \mathcal{B}(p)$ i.i.d.

Et alors ?

La loi des grands nombres nous dit que la valeur observée de

$$\overline{X}_n = \frac{1}{n}\Big(X_1 + X_2 + \cdots + X_n\Big)$$

devrait être raisonnablement proche de p.

Mais... si on recommençait aujourd'hui, on aurait une valeur différente.

Comment conclure quoi que ce soit en présence de hasard ?

Reste que pour l'instant, c'est notre meilleure *estimation* de p.

Ceci dit...

Si $X_i \sim \mathcal{B}(p)$, alors $\sum X_i \sim \mathcal{B}(n, p)$

espérance np, variance npq avec $q = 1 - p$

$$\implies \overline{X}_n = \frac{1}{n} \sum X_i$$

espérance p, variance $\dfrac{pq}{n}$

approximativement $\mathcal{N}\left(p, \underbrace{\sqrt{\frac{pq}{n}}}_{\sigma_n}\right)$ par TCL

Par exemple, on sait que $\overline{X}_n$ a 95 % de chances de tomber dans l'intervalle

$$\left[p - 2\,\sigma_n,\ p + 2\,\sigma_n\right].$$

En d'autres termes,

$$\begin{aligned} 0{,}95 &= \mathbb{P}[p - 2\,\sigma_n \leq \overline{X}_n \leq p + 2\,\sigma_n] \\ &= \mathbb{P}[\overline{X}_n - 2\,\sigma_n \leq p \leq \overline{X}_n + 2\,\sigma_n] \end{aligned}$$

C'est-à-dire : l'*intervalle aléatoire*

$$\left[\overline{X}_n - 2\,\sigma_n, \overline{X}_n + 2\,\sigma_n\right]$$

a 95 % de chances de contenir p !

Formalisons

Définition

Un **estimateur** est une variable aléatoire Θ_n dérivée d'un échantillon i.i.d. $(X_1, X_2, \dots, X_n)$ servant à estimer un paramètre θ de la loi des X_i.

Cet estimateur est dit **convergent** si

$$\lim_{n\to\infty} \Theta_n = \theta \quad \text{(presque sûrement).}$$

Il est **sans biais** si

$$\mathbb{E}[\Theta_n] = \theta \quad \text{pour tout } n.$$

Exemple vu et revu

$$\overline{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$$

est un estimateur de μ

- convergent (loi des grands nombres)
- sans biais (propriétés de $\mathbb{E}$).

Intervalle de confiance pour l'espérance

Pour n assez grand, on peut considérer que $\overline{X}_n \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$.

Si z_α désigne un nombre pour lequel

$$\mathbb{P}[-z_\alpha \leq Z \leq z_\alpha] = 1 - \alpha \quad \text{pour } Z \sim \mathcal{N}(0,1)$$

alors

$$\boxed{I_\alpha = \left[\overline{X}_n - z_\alpha \frac{\sigma}{\sqrt{n}}, \overline{X}_n + z_\alpha \frac{\sigma}{\sqrt{n}}\right]}$$

est un **intervalle de confiance** de niveau $1 - \alpha$ pour μ (exemple avec nos données).

Et la variance ?

$$S_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

semble une bonne idée

Il est bien convergent vers σ^2.

Petit problème : les n termes ne sont pas indépendants...

Proposition

$$\mathbb{E}\left[S_n^2\right] = \frac{n-1}{n}\sigma^2$$

Estimateur non biaisé de la variance

Vaut mieux donc préférer à S_n^2 la variation suivante :

$$\widetilde{S_n^2} := \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n\right)^2$$

qui a $\mathbb{E}\left[\widetilde{S_n^2}\right] = \sigma^2$.

Attention : cela ne signifie **PAS** que S_n est un estimateur sans biais de l'écart-type !

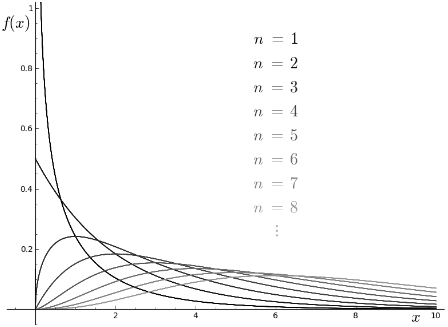
Rappel : sauf rares exceptions,

$$\mathbb{E}\left[\sqrt{X}\right] \neq \sqrt{\mathbb{E}[X]}$$

Loi de l'estimateur de la variance

Fait : si $X_i \sim \mathcal{N}(\mu, \sigma)$,

$$\frac{\widetilde{S_n^2}}{\sigma^2/(n-1)} \sim \chi^2_{n-1} \quad \text{loi du } \chi^2 \text{ à } n \text{ degrés de liberté}$$



En pratique

Notre intervalle de confiance pour μ

$$I_\alpha = \left[\overline{X}_n - z_\alpha\,\sigma,\ \overline{X}_n + z_\alpha\,\sigma\right]$$

supposait σ connu, dans les faits on doit l'estimer...

Pour un « grand » échantillon ($n > 30$) :

ça ne pose pas de problème de remplacer σ par $\widetilde{S_n}$.

(Pour un petit, on doit utiliser plutôt les quantiles d'une loi de Student)

Au menu aujourd'hui

Théorèmes limites

Estimation paramétrique

Tests d'hypothèse

Dans la vraie vie

On se pose des questions sur un modèle probabiliste pour prendre des décisions :

- ce dé est-il équilibré ?
- ce courriel est-il indésirable ?
- ce médicament est-il efficace ?
- cette machine est-elle déréglée ?
- ce candidat sera-t-il élu ?
- que faire face à ce risque ?
- combien rapportera ce placement ?
- cette mesure a-t-elle été efficace ?

⋮

Test d'hypothèse

Principe général : on tente d'*invalider* un modèle grâce à des observations.

- H_0 : **hypothèse nulle** décrivant un modèle probabiliste prédictif
- H_1 : **hypothèse alternative**

Si les observations effectuées sont *trop* improbables sous l'hypothèse H_0, on rejette cette hypothèse en faveur de H_1.

La déviation observée à H_0 est alors dite **statistiquement significative**.

Exemple : lancer de pièce

- H_0 : la pièce est équilibrée
- H_1 : pile est favorisé

Soit X le nombre de $\mathtt{P}$ en 10 lancers.

On juge qu'une observation avec $\mathbb{P} \leq 5\,\%$ remettrait en cause H_0.

Or, sous H_0, $X \sim \mathcal{B}(10, \frac{1}{2})$ et

$$\mathbb{P}[X \geq 9 \mid H_0] = \frac{10+1}{2^{10}} \approx 1{,}07\,\%$$

Si on observe $X \geq 9$, on pourra donc rejeter H_0 au seuil de signification $\alpha = 5\,\%$

Fonctionnement

- On choisit un **seuil de signification** α (souvent 5 % ou 1 %)
- On sélectionne une statistique T dont on connait la loi *sous* H_0
- On calcule la probabilité p que T prenne, *sous* H_0, une valeur aussi extrême que celle observée
- Si $p < \alpha$, on rejette H_0 en faveur de H_1 : la différence observée est **statistiquement significative**
- Si $p \geq \alpha$, on juge que les données ne sont pas suffisantes pour remettre en cause H_0 (status quo)

Attention

Il faut choisir le seuil de signification α **avant** de voir les données !

Et se méfier de la prolifération de tests. . .

Deux sortes d'erreurs possibles :

- rejeter H_0 alors qu'elle est vraie : se produit avec probabilité α prescrite
- accepter H_0 alors que H_1 est vraie : se produit avec une probabilité β

On appelle aussi $1 - \beta$ la **puissance** du test

Suite (et fin) la semaine prochaine