Probabilités, modélisation et statistique

Chapitre 5 - Statistique inférentielle

Raphaël Benerradi

Contenu pédagogique : Gwladys Toulemonde, Chloé Serre-Combe et Raphaël Benerradi

Polytech Montpellier - DevOps3 - Semestre 6

Année 2025-2026

Echantillon

Définition : échantillon

Une **échantillon** de taille n d'une variable aléatoire X est un n-uplet variables aléatoires $(X_1, \ldots, X_n)$ indépendantes et identiquement distribuées.
X est appelée **variable parente** de l'échantillon.

Remarque : En pratique, cela correspond à sélectionner de façon indépendante n individus a priori identiques (à des variations aléatoires près) parmi une population.

Remarque : Après une expérience, on recueille un jeu de données constitué des observations $(x_1, \ldots, x_n)$ qui correspondent aux réalisations de l'échantillon aléatoire $(X_1, \ldots X_n)$.

Statistique

Définition : statistique

Soit $(X_1, \ldots, X_n)$ un échantillon de taille n.
Une **statistique** est une variable aléatoire qui est fonction des variables aléatoires de l'échantillon.

$$S = g(X_1, \ldots, X_n), \qquad \text{avec } g : \mathbb{R}^n \to \mathbb{R}$$

Exemples :

- La moyenne empirique $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ est une statistique.
- Le maximum $\max(X_1, \ldots, X_n)$ et le p-ième percentile sont des statistiques.
- Le premier individu X_1 est une statistique.
- La variable aléatoire constante égale à 0 est une statistique (bien que très peu intéressante).

Estimateur et estimation

Définition : estimateur

Un **estimateur** est une statistique qui vise à fournir une estimation d'un paramètre θ de la loi parente d'un échantillon.

On note généralement un estimateur par un chapeau : $\hat{\theta}_n = g(X_1, \ldots, X_n)$.

Définition : estimation

Une **estimation** est la réalisation d'un estimateur pour un échantillon donné.

On la note parfois (abusivement) avec la notation de l'estimateur $(\hat{\theta}_n)$, mais cette fois-ci c'est bien une valeur numérique en non une variable aléatoire : $\hat{\theta}_n = g(x_1, \ldots, x_n) = g(X_1(\omega), \ldots, X_n(\omega))$ étant donné l'éventualité observée ω.

Remarque : A chaque fois qu'on réalise une expérience, l'estimation peut varier en fonction de l'échantillon observé. On comprend donc bien que l'estimateur est une variable aléatoire avec une loi de probabilité associée, là où l'estimation est un nombre réel.

Convergence

Définition : convergence d'un estimateur

Un estimateur $\hat{\theta}_n$ est dit **convergent** pour le paramètre θ s'il converge en probabilité quand la taille de l'échantillon n tend vers l'infini :

$$\hat{\theta}_n \xrightarrow[n\to\infty]{P} \theta$$

Biais et variance

Définition : biais

Le **biais** d'un estimateur $\hat{\theta}$ du paramètre θ est défini par :

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

Un estimateur est dit **sans biais** si $\text{Bias}(\hat{\theta}) = 0$.

Remarque : Un "bon" estimateur devrait avoir un biais faible, voire être sans biais.

Remarque : $\hat{\theta}$ étant une variable aléatoire, on peut tout à fait parler de sa **variance** $\mathbb{V}(\hat{\theta})$. On souhaite également que cette variance soit faible.

Erreur quadratique moyenne

Définition : erreur quadratique moyenne

L'**erreur quadratique moyenne** d'un estimateur $\hat{\theta}$ du paramètre θ est définie par :

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

L'erreur quadratique moyenne est souvent notée MSE pour *"Mean Squared Error"*.

Remarque : On peut montrer que

$$\text{MSE}(\hat{\theta}) = \mathbb{V}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$$

Ainsi, minimiser l'erreur quadratique moyenne revient à minimiser la somme de la variance et du biais de l'estimateur.

Cette décomposition illustre le compromis (ou dilemme) entre biais et variance dans la construction d'un estimateur. C'est un point important en apprentissage automatique (compromis entre sous-apprentissage et sur-apprentissage).

Qualités d'un estimateur

Plusieurs qualités sont attendues pour un "bon" estimateur :

- Convergence : quand la taille de l'échantillon augmente, on aimerait que l'estimateur converge en probabilité vers la vraie valeur du paramètre.
- Biais : on préferera un estimateur sans biais (i.e. $\text{Bias}(\hat{\theta}) = 0$) ou autant que possible avec un biais faible.
- Erreur quadratique moyenne : on cherchera un estimateur avec une erreur quadratique moyenne faible. Cela permet d'avoir à la fois un biais et une variance faibles.
- Robustesse : un bon estimateur devrait être robuste face aux variations de l'échantillon (ex : la médiane est plus robuste que la moyenne face aux valeurs aberrantes).

Estimateur de la moyenne empirique

Soit $(X_1, \dots X_n)$ un échantillon de variable parente X admettant un moment d'ordre 2. On veut estimer le paramètre $\theta = \mathbb{E}(X)$.

On choisit la **moyenne empirique** $\hat{\theta}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ pour estimer l'espérance de X.

Propriétés de la moyenne empirique

- La moyenne empirique est un estimateur sans biais
  $\mathbb{E}(\hat{\theta}_n) = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \times n\,\mathbb{E}[X] = \mathbb{E}[X] = \theta.$
- La variance de la moyenne empirique vaut :
  $\mathbb{V}(\hat{\theta}_n) = \mathbb{V}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{V}(X_i) = \frac{\mathbb{V}(X)}{n}.$
- Ainsi, $\text{MSE}(\hat{\theta}_n) = \mathbb{V}(\hat{\theta}_n) + \text{Bias}(\hat{\theta}_n)^2 = \frac{\mathbb{V}(X)}{n}.$
- La moyenne empirique est un estimateur convergent de $\theta = \mathbb{E}(X)$.
  En effet, la loi faible des grands nombres donne la convergence en probabilité (et même la convergence L^2).
  Le théorème central limite donne aussi une approximation de la distribution de $\hat{\theta}_n$ pour n grand : $\hat{\theta}_n \underset{n\to\infty}{\sim} \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$.

Exemple 1 : estimation d'une moyenne

On souhaite estimer la moyenne des émissions de CO_2-eq d'une population.

Pour cela, on tire n individus parmi la population (de façon indépendantes et uniforme). On a alors un échantillon de taille n qu'on suppose de variable parente $X \sim \mathcal{N}(\mu, \sigma^2)$, où μ est la moyenne des émissions de CO_2-eq dans la population, et σ^2 la variance. On peut utiliser la moyenne empirique pour estimer la moyenne μ :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

On sait maintenant que cet estimateur est convergent (en norme quadratique et en probabilité), sans biais, et de variance $\frac{\sigma^2}{n}$ (même si on ne connaît pas σ^2).
Ici on peut aussi dire que $\hat{\mu}$ suit une loi normale (comme somme de v.a. gaussiennes indépendantes).

Exemple 2 : estimation d'une proportion

On souhaite estimer la proportion de personnes végétariennes au sein d'une population.

Pour cela, on tire n individus parmi la population (de façon indépendantes et uniforme). On a alors un échantillon de taille n qu'on suppose de variable parente $X \sim \mathcal{B}(p)$, où p est la probabilité qu'un individu soit végétarien. On peut utiliser la moyenne empirique pour estimer cette proportion p :

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

On sait maintenant que cet estimateur est convergent (en norme quadratique et en probabilité), sans biais, et de variance $\frac{p(1-p)}{n}$ (même si on ne connaît pas p).

Estimateur de la variance empirique

Soit $(X_1, \dots X_n)$ un échantillon de variable parente X admettant un moment d'ordre 2. On veut estimer le paramètre $\theta = \mathbb{V}(X)$.

On va essayer l'estimateur de la **variance empirique**
$\hat{\theta}_n = S_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2$ pour estimer la variance de X.

Remarque :

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

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

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

Propriétés de la variance empirique

$$\mathbb{E}[\overline{X}_n^2] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right] = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[X_i X_j] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}[X_i^2] + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}[X_i]\mathbb{E}[X_j]$$

$$= \frac{1}{n^2}\, n\,\mathbb{E}[X^2] + \frac{1}{n^2}\, n(n-1)\,\mathbb{E}[X]^2 = \frac{\mathbb{E}[X^2]}{n} + \frac{(n-1)\,\mathbb{E}[X]^2}{n}$$

$$\mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] = \frac{1}{n}\, n\,\mathbb{E}[X^2] = \mathbb{E}[X^2]$$

Ainsi,

$$\mathbb{E}[S_n^2] = \mathbb{E}\left[\frac{1}{n}\left(\sum_{i=1}^{n} X_i^2\right) - \overline{X}_n^2\right] = \mathbb{E}[X^2] - \frac{\mathbb{E}[X^2]}{n} - \frac{(n-1)\,\mathbb{E}[X]^2}{n}$$

$$= \frac{n-1}{n}\left(\mathbb{E}[X^2] - \mathbb{E}[X]^2\right)$$

$$= \frac{n-1}{n}\mathbb{V}(X)$$

Propriétés de la variance empirique

- La variance empirique est un estimateur biaisé de la variance : $\mathbb{E}[S_n^2] = \frac{n-1}{n}\mathbb{V}(X)$.
- On préfèrera alors la **variance empirique corrigée** (ou variance empirique non biaisée) : $\widetilde{S}_n^2 = \frac{n}{n-1}S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$.

Idées

Quand on fait une estimation pour un paramètre θ, on aimerait avoir une idée de la "confiance" qu'on peut avoir dans cette estimation.

En particulier quand on a un seul échantillon, on a une seule valeur pour l'estimation de θ, et donc assez peu d'information sur la confiance qu'on peut avoir dans cette estimation.

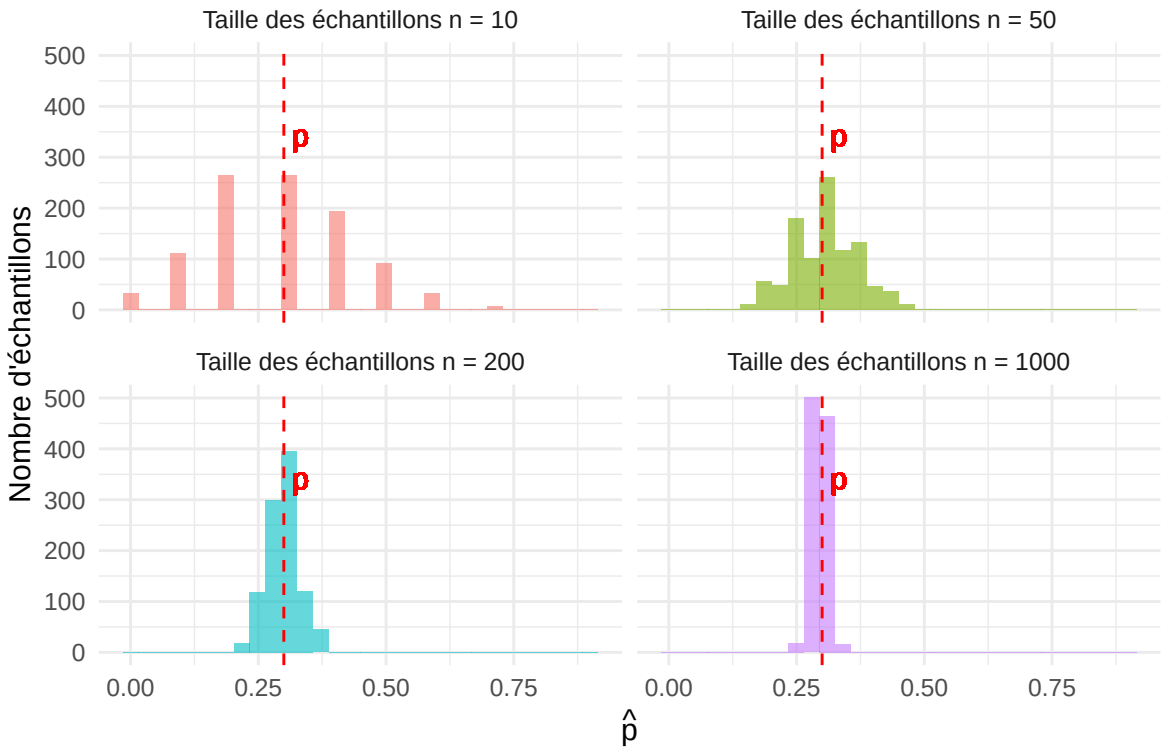
Ainsi, au lieu de donner une seule valeur $\hat{\theta}$ comme estimation, on peut :

- quand c'est possible, réchantillonner plusieurs fois pour estimer la loi de $\hat{\theta}$ (ex : méthode du bootstrap),
- construire un intervalle $[L, U]$ (qui dépend de l'échantillon) dans lequel la vraie valeur du paramètre θ a une forte probabilité de se trouver.

Estimer empiriquement la loi de $\hat{\theta}$

Exemple : On considère une population de 50000 individus suivant une loi de Bernoulli de paramètre $p = 0.3$. Pour une taille d'échantillon fixée ($n = 10, 50, 200,$ ou 1000), on peut échantillonner (tirer n individus) plusieurs fois, par exemple 500 fois. Pour chaque échantillon on peut estimer $\hat{p}$ avec la moyenne empirique.

On observe alors une distribution empirique de $\hat{p}$:



Intervalle de confiance

Définition : intervalle de confiance et niveau de confiance

Un **intervalle de confiance** pour un paramètre θ est un intervalle aléatoire $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$ construit à partir d'un échantillon $(X_1, \ldots, X_n)$ tel que :

$$\mathbb{P}\left(L(X_1, \ldots, X_n) \leq \theta \leq U(X_1, \ldots, X_n)\right) = 1 - \alpha$$

où $1 - \alpha$ est le **niveau de confiance** (souvent 0.95 ou 0.99).
α est le **risque** (i.e. ici la probabilité) de ne pas inclure la vraie valeur du paramètre dans l'intervalle $[L(X_1, \ldots, X_n), U(X_1, \ldots, X_n)]$.

Remarque : On s'attend à ce que :

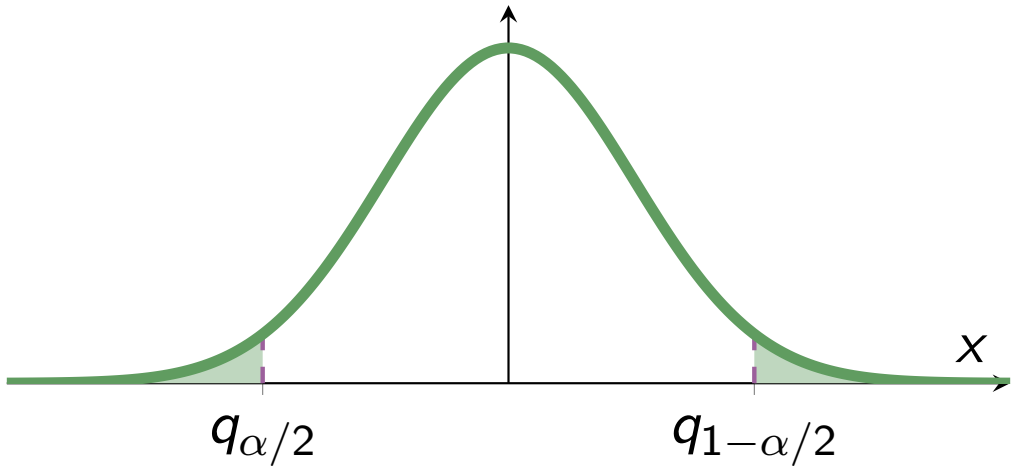
- à taille d'échantillon n fixée, quand on augmente le niveau de confiance $1 - \alpha$, l'intervalle de confiance $[L, U]$ s'élargisse.
- à niveau de confiance $1 - \alpha$ fixé, quand on augmente la taille d'échantillon n, l'intervalle de confiance $[L, U]$ se resserre.

Exemple 1 : estimation d'une moyenne

Soit $(X_i)_{i\in[\![1,n]\!]}$ i.i.d. de loi $\mathcal{N}(\mu,\sigma^2)$, et supposons la variance σ^2 connue.

On considère l'estimateur de la moyenne : $\hat{\mu} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. On a vu que $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ comme somme de v.a. gaussiennes indépendantes.

Soit $Z = \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$.
On pense naturellement à construire un intervalle de confiance pour μ à partir de Z, basé sur les quantiles de la loi normale centrée réduite.



Exemple 1 : estimation d'une moyenne

Soit $\alpha \in [0,1]$ le risque. Notons $q_{1-\alpha/2} \geq 0$ le quantile d'ordre $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$. La loi normale étant symétrique, on a $q_{\alpha/2} = -q_{1-\alpha/2}$ le quantile d'ordre $\alpha/2$.

$$\mathbb{P}\left(q_{\alpha/2} \leq Z \leq q_{1-\alpha/2}\right) = 1-\alpha$$

$$\mathbb{P}\left(-q_{1-\alpha/2} \leq \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq q_{1-\alpha/2}\right) = 1-\alpha$$

$$\mathbb{P}\left(-q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{X}_n - \mu \leq q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$$\mathbb{P}\left(\overline{X}_n - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X}_n + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Ainsi, un intervalle de confiance pour μ au niveau de confiance $1-\alpha$ est donné par :

$$\left[\overline{X}_n - q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X}_n + q_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

Exemple 2 : estimation d'une proportion

Soit $(X_i)_{i\in[\![1,n]\!]}$ i.i.d. de loi $\mathcal{B}(p)$.

On considère l'estimateur de la proportion : $\hat{p} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.
Le théorème central limite donne une approximation de la distribution de $\overline{X}_n$ pour n grand : $Z_n = \frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,1)$.

Remarque : La convergence en loi donne aussi la convergence des quantiles.

Exemple 2 : estimation d'une proportion

Avec $q_{\alpha/2}$ et $q_{1-\alpha/2}$ les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ de la loi $\mathcal{N}(0,1)$, on a donc :

$$\lim_{n\to\infty} \mathbb{P}\left(q_{\alpha/2} \leq Z_n \leq q_{1-\alpha/2}\right) = 1-\alpha$$

$$\lim_{n\to\infty} \mathbb{P}\left(-q_{1-\alpha/2} \leq \frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}} \leq q_{1-\alpha/2}\right) = 1-\alpha$$

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{X}_n - q_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \leq p \leq \overline{X}_n + q_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1-\alpha$$

On peut montrer que $\forall u \in [0,1]$, $u(1-u) \leq \frac{1}{4}$. D'où :

$$\lim_{n\to\infty} \mathbb{P}\left(\overline{X}_n - \frac{1}{2\sqrt{n}} \leq p \leq \overline{X}_n + \frac{1}{2\sqrt{n}}\right) \geq 1-\alpha$$