Rappels statistiques

Hugo Harari-Kermadec

M1 APE - Techniques quantitatives

2020

1 Introduction

Definition (Echantillon)

C'est l'ensemble des individus sur lesquels on a des informations.

Definition (Population)

La population est l'ensemble des individus qui pourraient être dans l'échantillon. Ce n'est pas "ce que nous voudrions étudier".

Example (Système universitaire)

Les 81 universités françaises, indicés par $i \in [|1; 81|]$. On peut par exemple mesurer le nb d'étudiants de l'établissement, h_i.

Definition (Echantillon i.i.d.)

$J_1, \ldots, J_n$ sont les indices de n individus tirés indépendamment et avec la même probabilité dans la population. Il faut donc tirer **avec remise**.

Example (Taux de femmes : sondage)

On tire 15 universités, c'est-à-dire des j, $J_1, J_2, \ldots, J_{10}$ Le nombre d'hommes dans chaque est déterministe : $H_1 = h_{J_1} \ldots$

Definition (Paramètre)

Un paramètre a est une caractéristique réelle (un nombre) de la distribution d'une variable dans la population.

Example (Nombre d'étudiants moyen dans les universités françaises)

Soit h, le nombre d'étudiants moyen dans la population étudiante en France :

$$h = \frac{1}{81} \sum_{j=1}^{81} h_j, \text{ où } h_j \text{ le nombre d'étudiants de l'université j}$$

Definition (Estimateur)

Un estimateur $\hat{h}_n$ du paramètre h est une fonction des données, qu'on veut proche de la vraie valeur du paramètre.

Example (Estimateur du nombre d'étudiants moyen dans la population : moyenne dans l'échantillon)

Si l'échantillon $h_1, \ldots, h_{15}$ est i.i.d. $\hat{h}_{15} = \frac{1}{15} \sum_{j=1}^{15} H_j$, converge par la loi des grands nombres vers l'espérance commune des H_j : $\mathbb{E}[H] = \frac{1}{81} \sum_{i=1}^{81} h_i$ car chaque H_j dans l'échantillon suit la même loi, qui lui donne une chance sur 81 d'être chacun des h_i. C'est la propriété de l'échantillon i.i.d.

L'estimateur est-il bon ?

Definition (Intervalle de confiance - CI)

$[a_n; b_n]$ est un intervalle de confiance à 95% pour a ssi

$$\mathbb{P}(h \in [a_n \ ; \ b_n]) = 0,95.$$

Le paramètre appartient à l'intervalle avec une probabilité 0,95.

L'estimateur est-il bon ?

Definition (Intervalle de confiance - CI)

$[a_n; b_n]$ est un intervalle de confiance à 95% pour a ssi

$$\mathbb{P}(h \in [a_n \ ; \ b_n]) = 0,95.$$

Le paramètre appartient à l'intervalle avec une probabilité 0,95.

Souvent, on n'a qu'une confiance asymptotique:

$$\mathbb{P}(h \in [a_n \ ; \ b_n]) \xrightarrow{n\to\infty} 0,95.$$

Example (Sondage électoral)

La popularité de Macron est estimée à 27%, avec $n = 1.000$, signifie qu'avec une probabilité 0,95 elle est dans [24% ; 30%]

CI: formule

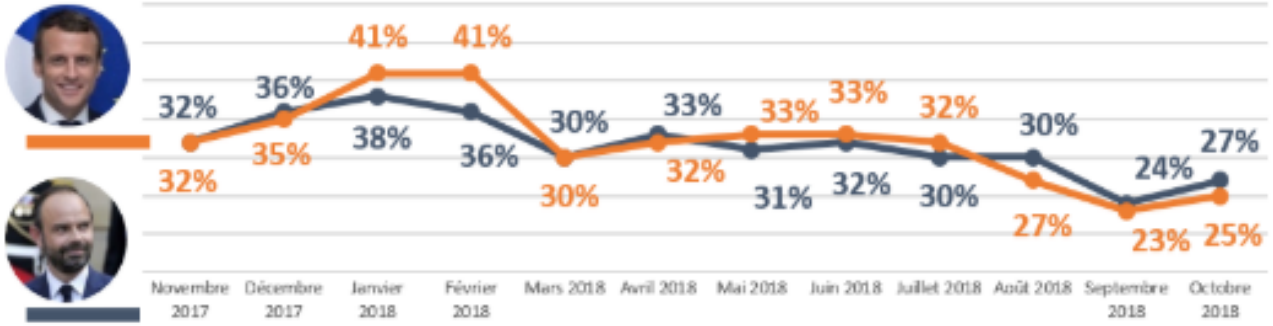
CI autour de la moyenne s'écrit :

$$[\hat{h}_n - q_{0.95}\frac{\sigma}{\sqrt{n}} \; ; \; \hat{h}_n + q_{0.95}\frac{\sigma}{\sqrt{n}}]$$

où $q_{0,95} \approx 2$ est le quantile à 95% d'une $\mathcal{N}(0,1)$,
et σ l'écart-type dans l'échantillon.

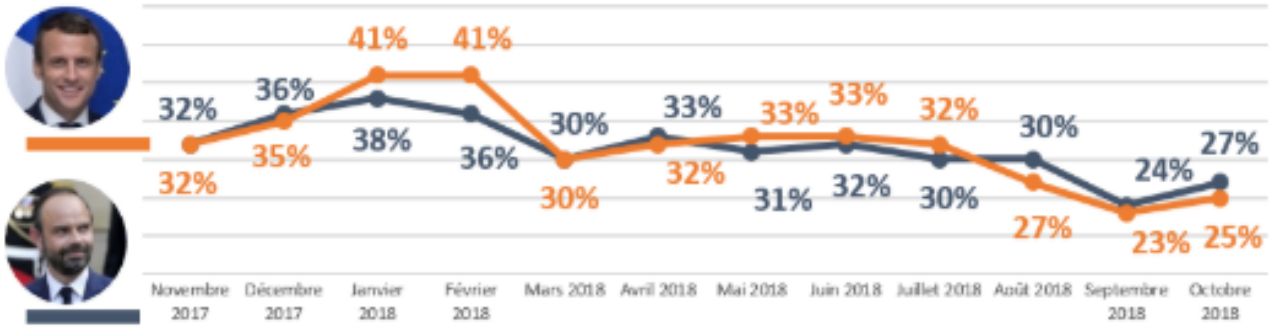
Exemple : popularité de Macron (YouGov)

La popularité de Macron, à 27% en août 2018 ($\Leftrightarrow$ [24%;30%]), est tombée en sept. à 23% puis remontée à 25% en oct. (1 006 individus)



Exemple : popularité de Macron (YouGov)

La popularité de Macron, à 27% en août 2018 ($\Leftrightarrow$ [24%;30%]), est tombée en sept. à 23% puis remontée à 25% en oct. (1 006 individus)



Les CI à 95% sont :

$$\left[23 - \frac{1/4}{\sqrt{1007}}\, q_{0,95} \;;\; 23 + \frac{1/4}{\sqrt{1007}}\, q_{0,95}\right] = [20 \;;\; 26]$$

$$\left[25 - \frac{1/4}{\sqrt{1007}}\, q_{0,95} \;;\; 25 + \frac{1/4}{\sqrt{1007}}\, q_{0,95}\right] = [22 \;;\; 28]$$

Une variation réduite, inférieure à 3 pts, n'est pas significative statistiquement. Par contre elle serait légalement significative pour une élection !

Definition (Test)

Soit h_0 une valeur possible pour le paramètre h. Si elle est dans le CI à 95%, alors l'hypothèse "nulle" H_0 : "$h = h_0$" est acceptée avec confiance 95%. Sinon, elle est rejetée.
La probabilité de rejeter H_0 à tort est notée α, elle vaut ici 5%.

Definition (Test)

Soit h_0 une valeur possible pour le paramètre h. Si elle est dans le CI à 95%, alors l'hypothèse "nulle" H_0 : "$h = h_0$" est acceptée avec confiance 95%. Sinon, elle est rejetée.
La probabilité de rejeter H_0 à tort est notée α, elle vaut ici 5%.

L'hypothèse d'une popularité constante entre sept. et oct. ($m = 0.23$) en acceptée, avec confiance 95%.

Principales caractéristiques des tests

	Accepte H_0	Rejette H_0
H_0 vraie	vrai positif $\mathbb{P}_{H_0} = 1 - \alpha$ Niveau du test	Faux négatif $\mathbb{P}_{H_0} = \alpha$ Erreur I
H_1 vraie (H_0 fausse)	Faux positif $\mathbb{P}_{H_1} = \beta$ Erreur II	Vrai négatif $\mathbb{P}_{H_1} = 1 - \beta$ Puissance du test

Plus précisément, l'hypothèse de stabilité est plus "faible" d'août à sept. ($27\% \to 23\%$) que de sept. à oct. ($23\% \to 25\%$)

Plus précisément, l'hypothèse de stabilité est plus "faible" d'août à sept. ($27\% \to 23\%$) que de sept. à oct. ($23\% \to 25\%$)
La p-value quantifie cette "force" :

Definition (p-value)

La p-value est l'erreur α maximale pour laquelle on accepte H_0.
Plus p-value est petite, plus l'hypohtèse est faible.

En juin, $p = 0.1$ et en septembre, $p = 0.05$.