Rappels statistiques

Hugo Harari-Kermadec

EOS - Économétrie

24 Septembre 2021

1 Introduction

- Lancez RStudio.
- Clic dans la console (en bas à gauche) et tapez 1+23.
- Posez `x<-3` et calculez x^2.

- Lancez RStudio.
- Clic dans la console (en bas à gauche) et tapez 1+23.
- Posez `x<-3` et calculez x^2.
- Posez `revenu <- c(873,1050,1401,4102,2350)` et calculez la moyenne, variance, écart-type, max, min, étendue, somme, longueur.

- Lancez RStudio.
- Clic dans la console (en bas à gauche) et tapez 1+23.
- Posez `x<-3` et calculez x^2.
- Posez `revenu <- c(873,1050,1401,4102,2350)` et calculez la moyenne, variance, écart-type, max, min, étendue, somme, longueur.
- et avec `age<-c(21,22,NA,51)` ?

- Lancez RStudio.
- Clic dans la console (en bas à gauche) et tapez 1+23.
- Posez `x<-3` et calculez x^2.
- Posez `revenu <- c(873,1050,1401,4102,2350)` et calculez la moyenne, variance, écart-type, max, min, étendue, somme, longueur.
- et avec `age<-c(21,22,NA,51)` ? Essayez `mean(age,na.rm=T)`

- Lancez RStudio.
- Clic dans la console (en bas à gauche) et tapez 1+23.
- Posez `x<-3` et calculez x^2.
- Posez `revenu <- c(873,1050,1401,4102,2350)` et calculez la moyenne, variance, écart-type, max, min, étendue, somme, longueur.
- et avec `age<-c(21,22,NA,51) ? Essayez mean(age,na.rm=T)`
- Posez `nationalité<-c("Française","Espagnole","Tunisienne")`

- Lancez RStudio.
- Clic dans la console (en bas à gauche) et tapez 1+23.
- Posez `x<-3` et calculez x^2.
- Posez `revenu <- c(873,1050,1401,4102,2350)` et calculez la moyenne, variance, écart-type, max, min, étendue, somme, longueur.
- et avec `age<-c(21,22,NA,51) ? Essayez mean(age,na.rm=T)`
- Posez `nationalité<-c("Française","Espagnole","Tunisienne")`

Toujours garder une copie

Créez un script en cliquant sur le + vert sous le menu *File*.
Sauvegardez le sous `basics.R` dans un dossier R1
Commentez avec ♯. On peut organiser le script avec des titres : ♯♯
On peut lancer le script sans passer par la console avec `Ctrl+Enter` ou `Cmd+Enter`. Si vous sélectionnez une partie du scripte, elle sera exécutée ; par défaut c'est l'ensemble du script.

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 (Revenu des salarié-es de l'ENS)

Soit les $N = 1\,000$ salarié-es de l'ENS Paris-Saclay. Pour chaque individu j, soit a_j son revenu. La population est indexée par $[|1; N|]$, avec les revenus $\{a_1, \ldots, a_N\}$.

- Téléchargez sur ecampus le fichier `population` et enregistrez le dans R1.
- Importez dans R avec le bouton *Import Dataset* (en haut à droite), “avec le format text (base)”
- copiez et collez le code produit dans basics.R

```
population <- read.csv("C:/.../R1/population.csv", sep=";")
```

Définissez le répertoire de travail en copiant ce chemin dans `setwd()`

Definition (Echantillon i.i.d.)

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

Example (Revenu des actifs : sondage)

On tire des individus: j's, $J_1, J_2, \ldots$ Le revenu est ensuite déterministe : $A_1 = a_{J_1}$

```
indexes<-sample(1:1000,20,replace=T)
```

Definition (Paramètre)

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

Example (Revenu moyen dans la population)

Soit a, le revenu moyen des salarié-es de l'ENS :

$$a = \frac{1}{N} \sum_{j=1}^{N} a_j, \text{ où } a_j \text{ revenu de l'individu j}$$

Soient a_M et a_F, les revenus moyens des hommes et des femmes :

$$a_M = \frac{1}{N_M} \sum_{j \text{ homme}}^{N} a_j, \qquad a_F = \frac{1}{N_F} \sum_{j \text{ femme}}^{N} a_j$$

Definition (Estimateur)

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

Example (Estimateur du revenu moyen dans la population : moyenne dans l'échantillon)

Si l'échantillon $a_1, \ldots, a_{100}$ est i.i.d. $\hat{a}_{100} = \frac{1}{100} \sum_{j=1}^{100} A_j$, converge par la loi des grands nombres vers l'espérance commune des A_j : $\mathbb{E}[A] = \frac{1}{N} \sum_1^N a_j$

On peut aussi estimer les revenus moyens par genre : with 45 males and 55 females

$\hat{a}_{M,45} = \frac{1}{45} \sum_{j \text{ homme}}^{n} a_j$, $\qquad$ $\hat{a}_{F,55} = \frac{1}{55} \sum_{j \text{ femme}}^{n} a_j$

L'estimateur est-il bon ?

Definition (Intervalle de confiance - CI)

$[l_n; u_n]$ est un intervalle de confiance à 95% pour a ssi

$$\mathbb{P}(a \in [l_n \ ; \ u_n]) = 0,95.$$

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

L'estimateur est-il bon ?

Definition (Intervalle de confiance - CI)

$[l_n; u_n]$ est un intervalle de confiance à 95% pour a ssi

$$\mathbb{P}(a \in [l_n \ ; \ u_n]) = 0,95.$$

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

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

$$\mathbb{P}(a \in [l_n \ ; \ u_n]) \xrightarrow{n \to \infty} 0,95.$$

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

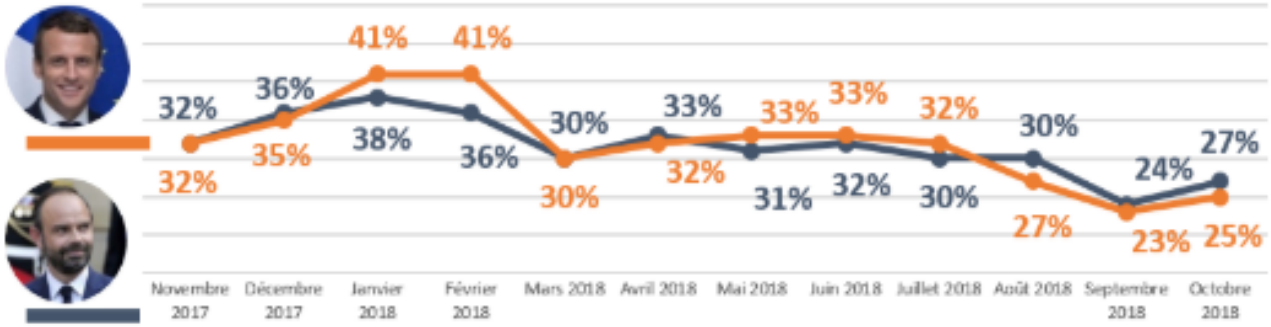
CI autour de la moyenne s'écrit :

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

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.

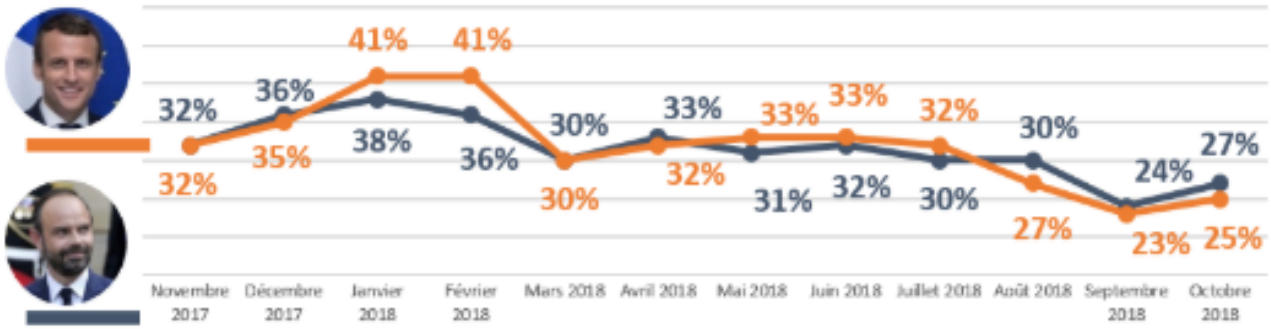
Exemple : popularité de Macron (YouGov)

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 a_0 une valeur possible pour le paramètre. Si elle est dans le CI à 95%, alors l'hypothèse "nulle" H_0 : "$a = a_0$" 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 a_0 une valeur possible pour le paramètre. Si elle est dans le CI à 95%, alors l'hypothèse "nulle" H_0 : "$a = a_0$" 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, $p = 0.05$.